Uncertainty-Aware Self-Supervised Cross-Modal SAR-Optical Matching Using EfficientDet and Xception

Peilong Huang[1], Yin Liu[2], Shengcheng Xie[2], Yimin An[2], Yifan Liang[2]
[1]State Grid Xinjiang Electric Power Co., LTD., Urumqi 830000, Xinjiang, China
[2]Electric Power Research Institute, State Grid Xinjiang Electric Power Co., LTD. Urumqi 830000, Xinjiang, China
E-mail: xj826712268@163.com

Keywords: Synthetic aperture radar, self-supervised learning, cross-modal matching, data fusion, data matching, machine learning, remote sensing, optical remote sensing, deep learning

Received: February 26, 2025

Cross-modal matching of Synthetic Aperture Radar (SAR) and optical satellite imagery is challenging due to their distinct imaging characteristics. We propose a deep learning framework integrating a dual encoder architecture, self-supervised contrastive learning, and uncertainty quantification for robust SAR-optical matching. The framework employs modality-specific encoders (EfficientDet for optical, Xception for SAR) with uncertainty modules capturing aleatoric and epistemic uncertainties, enhanced by self-supervised contrastive and rotation prediction tasks. Evaluated on the SEN12MS dataset, our method achieves a Maximum Mean Accuracy (MMA) of 0.145 at a 1-pixel threshold and 1298.3 average matched pairs per image (aNM), improving MMA by 20.8% over the state-of-the-art transformer-based method. Our uncertainty quantification yields an Expected Calibration Error (ECE) of 0.09, ensuring reliable confidence estimates. Ablation studies confirm the efficacy of our components, with computational efficiency improved by 40% faster convergence during supervised fine-tuning due to self-supervised pre-training. The method excels across diverse scenarios, including seasonal changes and varied land cover types, advancing SAR-optical matching for applications like change detection and disaster response.

Povzetek: Predstavljen je samonadzorovan model z negotovostjo za ujemanje SAR in optičnih slik, temelječ na ločenih enkoderjih in metodah za kvantifikacijo zanesljivosti ujemanja v daljinskem zaznavanju.

1 Introduction

The integration and matching of multi-modal remote sensing data, particularly between Synthetic Aperture Radar (SAR) and optical imagery, has emerged as a critical challenge in Earth observation applications [23, 17]. This growing importance is driven by the unprecedented availability of complementary data from missions such as Sentinel-1 and Sentinel-2, coupled with increasing demands for reliable Earth observation in various applications [9]. While optical sensors excel in providing rich spectral information under favorable conditions, SAR systems offer unique advantages through their weather-independent, day-and-night imaging capabilities [22]. The fundamental challenge in SAR-optical matching stems from the inherent differences in their imaging mechanisms and characteristics [20]. SAR imagery is characterized by speckle noise and distinctive geometric distortions resulting from its side-looking acquisition geometry, while optical imagery is subject to atmospheric conditions and illumination variations [27]. These differences manifest in several ways:

First, SAR imagery exhibits unique scattering properties where the signal return is strongly influenced by surface roughness and material properties, creating patterns that often have no direct correspondence in optical imagery [26]. Second, geometric distortions in SAR, including layover and foreshortening effects, are particularly pronounced in urban environments with complex 3D structures [12]. Third, the seasonal variations that significantly affect optical imagery may have minimal impact on SAR data, creating temporal matching challenges [21]. Traditional approaches to this matching problem have relied heavily on hand-crafted features and geometric constraints [17]. However, these methods often struggle to handle the complex, non-linear relationships between SAR and optical image characteristics [9]. Recent advances in deep learning have shown promising results in addressing these challenges, particularly through the development of specialized architectures for cross-modal feature learning [8]. Nevertheless, several critical limitations remain, such as, most existing approaches treat SAR and optical data uniformly, without fully accounting for their distinct characteristics and noise patterns [5]. The scarcity of accurately matched training data, especially in complex urban environments, also continues to be a significant bottleneck [6] and current methods often lack robust uncertainty quantification, making it difficult to assess the reliability of matching results in operational scenarios [11].

To address these challenges, this research aims to improve SAR-optical matching through self-supervised learn-

ing and uncertainty-aware deep learning models, through the following research questions:

1. How does uncertainty quantification improve SAR-optical image matching reliability?
2. Can self-supervised learning reduce the dependence on manually labeled data?
3. How does the proposed approach compare to state-of-the-art methods in various environmental conditions?

These objectives guide our development of a novel framework that leverages a dual encoder architecture, self-supervised learning, and uncertainty quantification to achieve robust and reliable SAR-optical matching. Our approach employs modified EfficientDet and Xception networks as modality-specific encoders, chosen for their proven effectiveness in handling complex visual tasks. EfficientDet's scalable feature extraction is adapted for optical imagery to capture multi-spectral details despite atmospheric variations, while Xception's depthwise separable convolutions are tailored for SAR to mitigate speckle noise and geometric distortions [25, 8]. These modifications ensure that each encoder is optimized for the unique imaging characteristics of its respective modality. The framework incorporates a self-supervised learning strategy with contrastive learning and rotation prediction tasks. Contrastive learning aligns SAR and optical features in a shared embedding space to address the domain gap, while rotation prediction enhances geometric invariance, overcoming limitations of prior methods that struggle with viewpoint differences and scarce labeled data [4]. An uncertainty-guided matching mechanism dynamically weighs feature similarities based on estimated confidence levels, integrating aleatoric and epistemic uncertainties. This mechanism improves matching accuracy by prioritizing reliable correspondences and enhances reliability by providing interpretable confidence estimates, crucial for mission-critical applications [11].

Our approach advances the state-of-the-art by integrating uncertainty quantification, leveraging self-supervised learning, and improving robustness across diverse scenarios. First, by incorporating modality-specific uncertainty quantification, we enable more reliable matching decisions while providing interpretable confidence measures. Second, our self-supervised learning strategy effectively leverages the abundance of unpaired SAR and optical data, reducing the dependence on manually matched training samples. Third, our framework demonstrates robust performance in various challenging scenarios, including complex urban environments and seasonal changes.

To quantify the performance of our method, we use metrics tailored for the SAR-optical matching problem. Specifically, we report a Maximum Mean Accuracy (MMA) of 0.145 at a 1-pixel threshold and an average of 1298.3 matched pairs per image (aNM). These results indicate that our approach not only achieves high precision in identifying correct matches (as reflected in MMA), but also produces a substantial number of reliable correspondences (as indicated by aNM), which is crucial for downstream applications like registration and change detection. These metrics reflect a 20.8% relative improvement over prior state-of-the-art methods, underscoring the practical and computational advantages of our framework.

The remainder of this paper is organized as follows: 2 reviews related work on SAR-optical matching, uncertainty estimation, and self-supervised learning. 3 presents our methodology, including the detailed architecture and training strategy. 4 describes our experimental setup and results, while 5 discusses the implications and limitations of our approach. Finally, 6 concludes the paper with a summary of our contributions.

2 Related work

2.1 SAR-optical image matching

The evolution of SAR-optical matching techniques can be broadly categorized into traditional feature-based methods and modern deep learning approaches. Traditional methods relied mainly on hand-made features and geometric constraints [17]. Notable examples include the Scale-Invariant Feature Transform (SIFT) adaptations [16] and Speeded Up Robust Features (SURF) variants [1], which were modified to handle the unique characteristics of SAR imagery. However, these approaches often struggled with the fundamental differences between SAR and optical imaging mechanisms [13]. Recent years have seen significant advances in deep learning-based approaches. Hughes et al. [8] demonstrated the effectiveness of pseudo-siamese architectures for patch-based matching, while Merkle et al. [18] explored conditional adversarial networks for bridging the domain gap between SAR and optical imagery. These approaches have shown promising results but often lack robust uncertainty quantification.

A significant breakthrough came with the release of large-scale datasets like SEN12MS [21], which enabled more comprehensive training and evaluation of deep learning models. This has led to various architectural innovations, including attention-based mechanisms [24] and transformer-based approaches [2] for feature matching.

To provide a clearer comparison of existing methods and highlight their shortcomings, we summarize key SAR-optical matching approaches in 1. The table compares traditional and deep learning methods in terms of datasets used, accuracy metrics, and main limitations. As shown, traditional methods like SIFT-based approaches suffer from poor performance on SAR imagery due to noise sensitivity, while deep learning methods, despite improved accuracy, often lack uncertainty quantification, require pre-aligned data, or are computationally intensive. These limitations underscore the need for our proposed method, which integrates uncertainty-aware feature extraction and self-supervised learning to achieve robust and reliable matching.

Table 1: Comparative summary of sar-optical image matching methods. note: accuracy metrics vary by pixel threshold, limiting direct comparison

Method	Dataset Used	Accuracy Metrics	Main Limitations Identified
SIFT-based [10]	Not specified	MMA: 2.94% (<2 px), 7.92% (<3 px), 13.01% (<4 px); Avg Error: 9.92 px	Poor performance on SAR; sensitive to noise and appearance variance.
RIFT [13]	Multiple multi-modal datasets	MMA: 8.6% (<2 px), 25.9% (<3 px), 53.6% (<4 px); Avg Error: 2.80 px	Computationally heavy; requires careful parameter selection.
Pseudo-Siamese CNN [7]	Automatically generated dataset	High patch matching accuracy	Limited to patches; lacks uncertainty modeling or spatial generalization.
CAM-Net + CAMM [3]	Not specified	Repeatability: 0.434; LE: 1.96 px; NN mAP: 0.3090; ACE: 7.15	High memory/computational demands; limited interpretability.
3-step CNN Framework [8]	Sentinel-1/2 urban scenes	Good spatial matching; no MMA value reported	Requires pre-aligned data; lacks real-world robustness assessment.
MOEFC [15]	Urban + rural SAR-optical pairs	RMSE: 0.645 px (rural), 0.489 px (urban)	Edge-based; sensitive to image structure and noise level.
Proposed Method (Efficient-Det+Xception)	SEN12MS	MMA@1px: 0.145; aNM: 1298.3; ECE: 0.09	High computational cost from uncertainty estimation; dataset-dependent.

2.2 Deep learning with SAR-optical data

The application of deep learning to SAR-optical data fusion has evolved significantly, particularly in addressing the unique challenges of each modality. Early work by Mou et al. [19] introduced CNN-based approaches for patch correspondence identification, achieving promising results but struggling with complex urban scenes. Recent advances have focused on more sophisticated architectures that can better handle the distinct characteristics of each modality [30]. The notable developments include the adaptation of modern CNN architectures for SAR-specific feature extraction [8], integration of attention mechanisms for better feature correlation [24], and development of multi-scale approaches for handling varying spatial resolutions [29].

2.3 Uncertainty in deep learning for remote sensing

Uncertainty quantification in remote sensing has gained significant attention, particularly for mission-critical applications. The literature distinguishes between aleatoric uncertainty (capturing noise inherent in observations) and epistemic uncertainty (representing model uncertainty) [11]. In the context of SAR-optical fusion, recent work has explored Bayesian neural networks and ensemble approaches for uncertainty-aware feature extraction [8]. The incorporation of uncertainty estimation has proven particularly valuable in several aspects, such as improved reliability assessment of matching results, better handling of challenging scenarios such as seasonal changes, and more robust performance in areas with significant temporal variations.

2.4 Self-supervised learning in remote sensing

Self-supervised learning has emerged as a powerful paradigm for leveraging unlabeled remote sensing data [4]. The abundance of unpaired SAR and optical imagery makes this particularly relevant for cross-modal matching. Recent work has shown the effectiveness of contrastive learning approaches in learning robust representations from satellite imagery [28]. The key developments in this area include contrastive learning strategies for cross-modal feature learning, rotation prediction tasks for geometry-aware feature extraction, and multi-task self-supervised frameworks that combine multiple pretext tasks. This self-supervised learning paradigm has proven particularly valuable in addressing the scarcity of labeled training data, a common challenge in remote sensing applications [14]. Recent work has demonstrated that pre-training on large amounts of unlabeled data can significantly improve the performance of downstream matching tasks [30].

3 Proposed methodology

Our proposed framework addresses the fundamental challenges of SAR-optical matching through a novel architecture that combines uncertainty-aware feature extraction with self-supervised learning technique using a dual en-

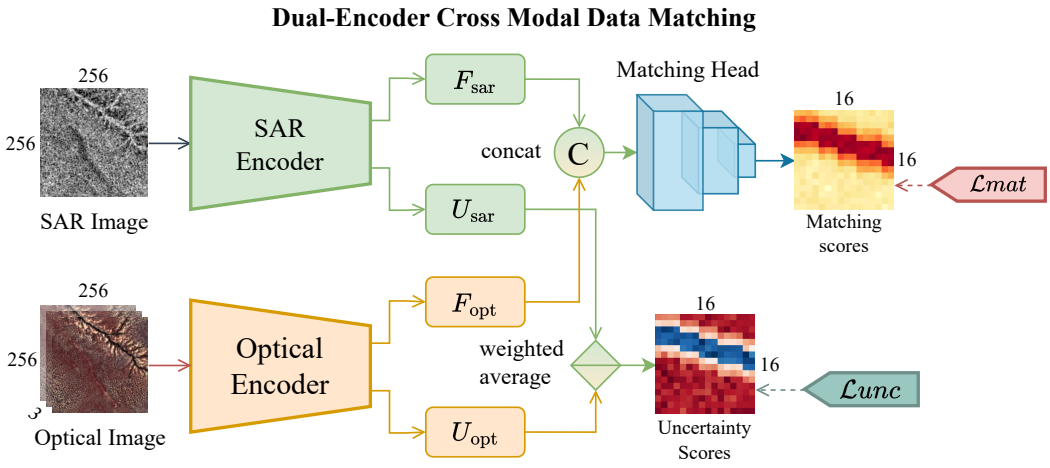

Figure 1: Overview of the proposed cross-modal matching architecture. The framework consists of parallel SAR and optical encoders (Xception and EfficientDet, respectively), each equipped with modality-specific uncertainty modules. The matching head combines features from both modalities while accounting for their respective uncertainties. $\mathcal{L}_{mat}$ and $\mathcal{L}_{unc}$ represent the matching loss and uncertainty loss, respectively.

coder architecture with modality-specific uncertainty modules and a matching mechanism. 1 illustrates the overall architecture of our dual encoder cross modal matching approach.

3.1 Dual encoder architecture

The foundation of our approach lies in recognizing that SAR and optical imagery require specialized processing streams due to their fundamentally different imaging characteristics [8]. Our dual encoder architecture implements this principle through modality-specific feature extractors. Figures 2 and 3 illustrate the full architectures of the SAR and optical encoders, respectively, including their uncertainty modules, which are detailed in 3.2.

3.1.1 SAR encoder

For SAR imagery, we adopt a modified Xception architecture optimized for the unique properties of radar backscatter. 2 shows the architecture of the SAR encoder where the key modifications include initial convolution layer (stride 2) adapted for single-channel SAR input followed by four modified Xception blocks with increasing channel dimensions ($32 \rightarrow 64 \rightarrow 128 \rightarrow 256 \rightarrow 512$), incorporating specialized separable convolutions with residual connections, followed by SAR-specific uncertainty module calibrated for speckle noise, geometric distortions, and radar backscatter characteristics. This uncertainty module enhances robust uncertainty quantification by capturing modality-specific noise patterns, addressing the limitations of existing methods that often lack such reliability measures [5], as detailed in 3.2.1. The SAR encoder, (E_{sar}) can be formally expressed as:

$$F_{\text{sar}}, U_{\text{sar}} = E_{\text{sar}}(X_{\text{sar}}; \theta_{\text{sar}}) \quad (1)$$

where F_{sar} represents the extracted SAR features, U_{sar} denotes the uncertainty estimates and θ_{sar} represents the parameters of the SAR encoder.

3.1.2 Optical encoder

The optical stream employs a modified EfficientDet architecture [25] shown in 3 optimized for multi-spectral imagery:

$$F_{opt}, U_{opt} = E_{opt}(X_{\text{opt}}; \theta_{\text{opt}}) \quad (2)$$

where F_{opt} represents the extracted optical features, U_{opt} denotes the uncertainty estimates and (θ_{opt}) represents the parameters of the optical encoder, (E_{opt}). The key architectural elements include an initial 7×7 convolution layer for processing RGB input, three feature extraction layers with increasing channels ($64 \rightarrow 128 \rightarrow 256 \rightarrow 512$), a Bi-directional Feature Pyramid Network (BiFPN) for multi-scale feature fusion, and an optical-specific uncertainty module accounting for atmospheric variations, illumination changes, and seasonal variations. This module enables robust uncertainty quantification by modeling variations inherent to optical imagery, overcoming the shortcomings of prior approaches that lack such capabilities [5], as elaborated in 3.2.2.

3.2 Uncertainty quantification

A key challenge in SAR-optical matching lies in understanding when and why our model might make incorrect matches. Our uncertainty quantification approach addresses this challenge by recognizing that SAR and optical images have fundamentally different characteristics that can lead to matching errors. Rather than treating both modalities the same way, we develop specialized uncer-

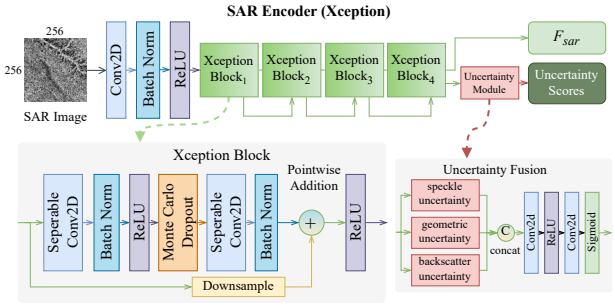

Figure 2: Architecture of the SAR-specific encoder, including the modified Xception blocks with an initial convolution layer (stride 2) and the uncertainty module (detailed in 3.2.1), which outputs both features (F_{sar}) and uncertainty scores.

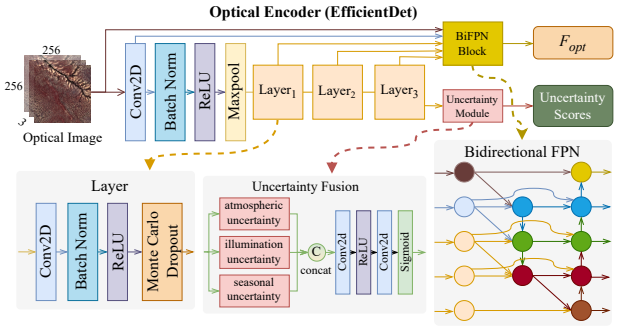

Figure 3: Architecture of the optical encoder, including the modified EfficientDet components with a 7×7 convolution layer, BiFPN, and the uncertainty module (detailed in 3.2.2), which outputs both features (F_{opt}) and uncertainty scores.

tainty modules that capture the unique challenges of each imaging type.

3.2.1 SAR-specific uncertainty considerations

For SAR imagery, we identify and address three primary sources of uncertainty that can affect matching reliability, as shown in 4. First, we consider speckle noise, a characteristic feature of SAR imagery that appears as a grainy texture pattern. To address this, we implement a specialized processing branch that uses depthwise convolutions to analyze each channel independently. This branch consists of two Conv2D, each followed by ReLU activation, to extract features sensitive to speckle noise patterns, increasing uncertainty in noisy regions. This approach allows the model to better understand how speckle patterns might influence feature reliability in different parts of the image. Secondly, we tackle the geometric distortions inherent in SAR imagery due to its side-looking acquisition geometry. Our solution employs orientation-aware convolutions that analyze features at multiple angles (0°, 45°, 90°and 135°). This mechanism uses four parallel Conv2D layers, one for each angle,

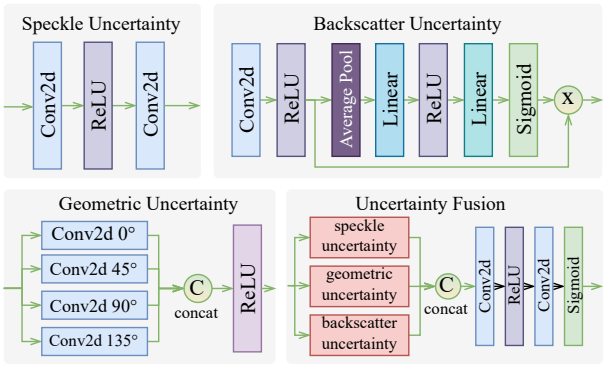

Figure 4: Architecture of the individual modules of the aleatoric uncertainty in SAR images.

whose outputs are concatenated and processed through a ReLU activation to capture geometric distortions across different orientations. This multi-orientation approach helps the model identify when geometric distortions might make matching unreliable. And lastly, we address backscatter intensity variations, which can create ambiguities in feature matching. We implement an attention mechanism that helps the model focus on stable backscatter patterns while expressing higher uncertainty in areas where backscatter variations might lead to unreliable matches. This attention mechanism applies a Conv2D layer, ReLU, global average pooling, two linear layers, ReLU, and a Sigmoid activation to generate attention weights, which are pointwise multiplied with the input features to emphasize stable backscatter regions. The outputs of these branches are combined in an uncertainty fusion module, which concatenates the three uncertainty maps, processes them through a Conv2D layer, ReLU activation, and a Sigmoid function to produce a final uncertainty map. These strategies address remote sensing challenges by ensuring robust matching across diverse land cover types, such as urban and rural areas, where speckle noise and geometric distortions vary significantly, enhancing reliability for applications like disaster monitoring [22].

3.2.2 Optical-specific uncertainty considerations

Likewise, for optical imagery, we focus on three different sources of uncertainty, as shown in 5. Atmospheric variations can significantly impact optical image quality and feature appearance. Our atmospheric uncertainty branch employs wide-area spatial attention mechanisms to identify regions where atmospheric effects might compromise matching reliability. This branch processes the input through a Conv2D layer, followed by ReLU activation and global average pooling, producing a spatial attention map. The map is pointwise multiplied with the input features to emphasize regions affected by atmospheric effects like clouds or haze, assigning higher uncertainty to those areas. This approach is particularly effective because atmospheric effects typically impact larger spatial regions coherently. Illumination conditions present another significant challenge in optical

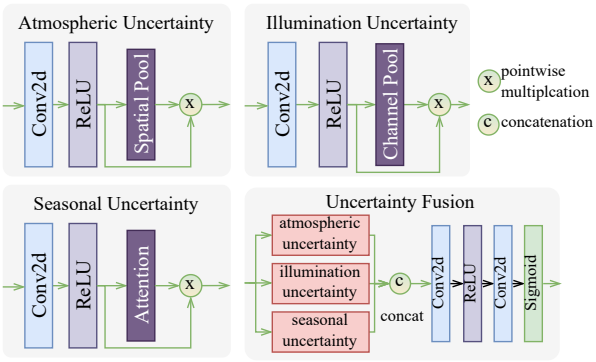

Figure 5: Architecture of the individual modules of the aleatoric uncertainty in optical images

imagery. We address this through a channel-wise attention mechanism that analyzes relationships between different spectral bands, helping identify when illumination conditions might make matching uncertain. This mechanism applies a Conv2D layer, ReLU activation, and global average pooling to generate channel-wise attention weights, which are pointwise multiplied with the input to highlight illumination inconsistencies across RGB bands. This is particularly important because SAR imagery is not affected by these illumination variations. Seasonal changes can create significant appearance differences in optical imagery. Our seasonal change uncertainty branch specifically focuses on identifying areas where temporal variations between image acquisitions might affect matching reliability. This branch uses a Conv2D layer, ReLU activation, and an attention mechanism (pointwise multiplication with learned weights) to detect seasonal variations like vegetation or snow cover changes, increasing uncertainty in affected regions. This helps the model express appropriate uncertainty in regions where seasonal changes might compromise matching accuracy. The outputs of these branches are combined in an uncertainty fusion module, which concatenates the three uncertainty maps, processes them through a Conv2D layer, ReLU activation, and a sigmoid function to produce a final uncertainty map. These mechanisms tackle remote sensing challenges by improving robustness to temporal variability and atmospheric effects, critical for consistent performance across seasons and geographic regions in applications like change detection [23].

3.2.3 Integrating multiple sources of uncertainty

Rather than treating these different sources of uncertainty independently, we recognize that they often interact in complex ways. Our approach combines the individual uncertainty estimates through an adaptive fusion mechanism that learns to weight different sources of uncertainty based on the specific characteristics of each image pair. This fusion approach proves particularly valuable in complex scenarios where multiple sources of uncertainty might be present. The effectiveness of this approach to uncertainty estimation is demonstrated in our experimental results, where we observe that the model not only achieves better matching accuracy but also provides well-calibrated confidence estimates. This is particularly important in practical applications where understanding the reliability of matches is crucial for downstream decision-making processes.

3.2.4 Model uncertainty through Monte Carlo dropout

While our modality-specific uncertainty modules capture uncertainties related to data characteristics, we also need to account for uncertainty in the model's predictions themselves. This is particularly important in cross-modal matching where the model might encounter scenarios different from its training data. We address this through Monte Carlo dropout, a technique that helps us estimate the model's prediction uncertainty by simulating an ensemble of slightly different models. The key insight behind this approach is that by randomly deactivating different parts of the network during inference (using dropout), we can obtain multiple predictions for the same input. The variation in these predictions gives us valuable information about the model's confidence. In regions where predictions are consistent across different dropout patterns, we can be more confident in our results. Conversely, high variation in predictions suggests uncertainty in the model's decision. We carefully position dropout layers throughout the network to capture uncertainty at different processing stages: 1) Early in the encoders (10% dropout rate) to capture uncertainty in initial feature extraction, 2) before uncertainty estimation (20% dropout rate) to ensure robust uncertainty predictions, and 3) in feature fusion (15% dropout rate) to capture uncertainty in the integration process.

By using 20 forward passes during inference, we obtain a reliable estimate of model uncertainty while maintaining practical computational efficiency. This approach proves particularly valuable in identifying challenging cases where the model might be encountering patterns significantly different from its training data.

3.3 Self-supervised pretraining

While the SEN12MS dataset provides co-registered images, we can significantly improve our model's performance by learning patterns from the SAR and optical images in a self-supervised manner. We achieve this through a self-supervised learning strategy that helps the model understand the relationship between modalities even without using explicit matching information from the dataset. 6 gives an overview of the self-supervised learning strategy.

3.3.1 Cross-modal understanding through contrastive learning

Our contrastive learning approach helps the model understand the fundamental relationships between SAR and optical representations of the same scene. The key idea is to

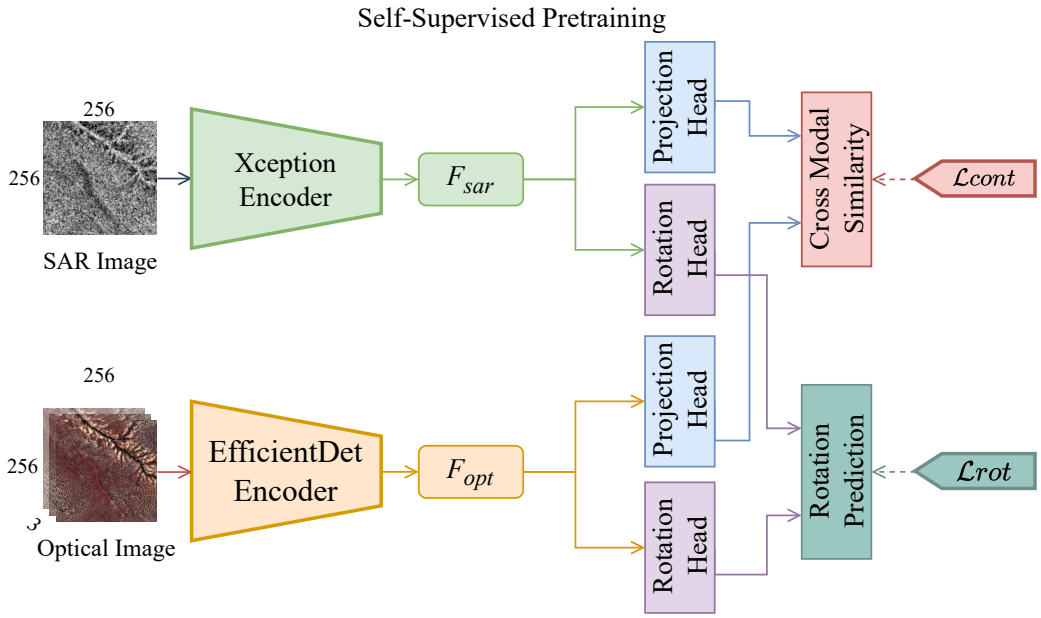

Figure 6: Overview of self-supervised learning strategy with contrastive learning and geometric understanding. $\mathcal{L}_{cont}$ and $\mathcal{L}_{rot}$ represent the contrastive loss and rotation loss, respectively.

teach the model to recognize when SAR and optical images represent the same location, even if they look very different due to their distinct imaging characteristics. We implement this through a specialized projection head that maps features from both modalities into a shared 128-dimensional space where meaningful comparisons can be made. The projection head is a 2-layer MLP with 512 and 128 units, using ReLU activation after the first layer, applied to both SAR and optical features (F_{sar} and F_{opt}). The cross-modal similarity is computed using cosine similarity in the shared space, defined as:

$$\text{sim}(z_{\text{sar}}, z_{\text{opt}}) = \frac{z_{\text{sar}} \cdot z_{\text{opt}}}{\|z_{\text{sar}}\| \|z_{\text{opt}}\|}, \tag{3}$$

where z_{sar} and z_{opt} are the projected features. The learning process encourages the model to bring representations of the same location closer together while pushing representations of different locations apart. A carefully chosen temperature parameter (0.07) helps maintain the right balance between positive and negative examples. The contrastive loss, $\mathcal{L}_{\text{cont}}$, is the InfoNCE loss, formulated as:

$$\mathcal{L}_{\text{cont}} = -\log \frac{\exp(\text{sim}(z_{\text{sar}}, z_{\text{opt}})/\tau)}{\sum_{k \neq \text{opt}} \exp(\text{sim}(z_{\text{sar}}, z_k)/\tau)}, \tag{4}$$

where $\tau = 0.07$, and the denominator sums over negative samples from the batch. 7 illustrates the effectiveness of our approach in learning a shared embedding space (We used Principal Component Analysis (PCA) to reduce the 128-dimensional embedding space to three dimensions for visualization) where meaningful cross-modal comparisons can be made despite the inherent differences between SAR and optical imagery.

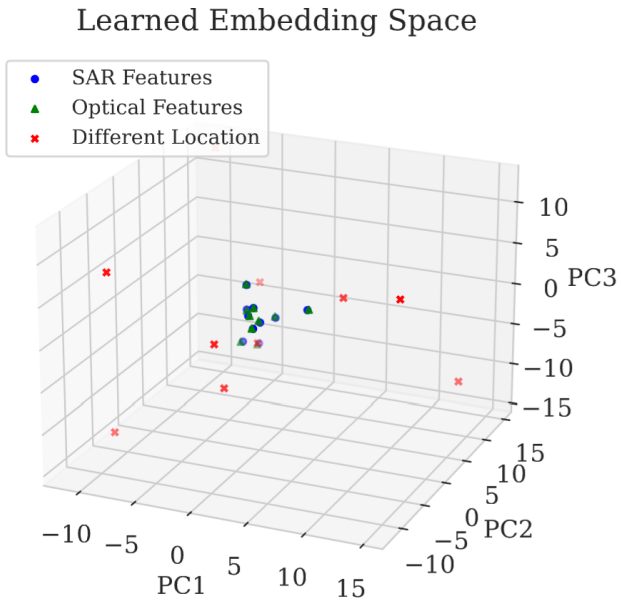

Figure 7: Visualization of features after applying contrastive learning with temperature $\tau = 0.07$

3.3.2 Geometric understanding through rotation prediction

While contrastive learning helps with feature similarity, we also need the model to understand geometric relationships between modalities. We achieve this through a rotation prediction task, where the model learns to identify the rotation applied to input images. This seemingly simple task actually helps the model develop a deeper understanding of ge-

Figure 8: Distribution of dataset samples across different categories

ometric structures that are preserved across modalities. The rotation head, applied to both SAR and optical features, is a 2-layer MLP with 256 and 4 units, followed by a softmax layer to predict four rotation angles ($0°$, $90°$, $180°$, $270°$). The rotation loss, $\mathcal{L}_{\text{rot}}$, is a cross-entropy loss, defined as:

$$\mathcal{L}_{\text{rot}} = -\sum_{c=1}^{4} y_c \log(\hat{y}_c), \tag{5}$$

where y_c is the true rotation label, and $\hat{y}_c$ is the predicted probability for class c. This loss is computed separately for SAR and optical images and averaged. By predicting rotations for both SAR and optical images, the model learns to identify geometric patterns that are consistent between modalities, even when the visual appearances are quite different. This geometric understanding is crucial for reliable cross-modal matching.

4 Experiments and results

4.1 Dataset and preprocessing

We conduct our experiments using the SEN12MS dataset [21], which provides an extensive collection of 180,662 co-registered SAR-optical image pairs from Sentinel-1 and Sentinel-2 satellites. The dataset's co-registration ensures accurate ground truth for evaluating matching reliability and uncertainty quantification. Its extensive coverage of unpaired SAR and optical data supports self-supervised learning, reducing reliance on labeled samples. Additionally, SEN12MS spans diverse geographical regions (e.g., urban, rural, forested areas), seasonal conditions (e.g., summer, winter), land cover types, and atmospheric variations (e.g., clear, cloudy), making it ideal for assessing robustness across challenging scenarios. For our experiments, we preprocess the patches by cropping and resizing to 256×256 pixels, normalizing pixel values to [0, 1], and applying data augmentation. We divide the dataset using a stratified sampling approach to maintain representative distribution across different categories. The resulting split consists of 126,463 pairs (70%) for training, 27,099 pairs (15%) for validation, and 27,100 pairs (15%) for testing. 8 shows the distribution of samples across different land cover types and seasonal conditions.

4.2 Implementation details

Our implementation uses PyTorch and is trained on a distributed system comprising 4 NVIDIA V100 GPUs. 2 summarizes the key training parameters for both the self-supervised pre-training and supervised fine-tuning phases. To facilitate replication, we provide additional details on the training setup. We use the Adam optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and a weight decay of $1e-4$. Data augmentation includes random rotations ($\pm 30°$), horizontal flips (50% probability), and Gaussian noise (standard deviation 0.01) applied to both SAR and optical inputs to enhance robustness. For the SAR encoder, the input resolution is 256×256 with a single channel, while the optical encoder processes RGB inputs at the same resolution. The contrastive learning projection head uses a 2-layer MLP with 512 and 128 units, and the rotation prediction task classifies four angles ($0°$, $90°$, $180°$, $270°$). The Monte Carlo dropout employs a 20-pass inference with dropout rates as specified in 3.2.4.

4.3 Evaluation metrics

To comprehensively assess the performance of our approach, we employ several complementary evaluation metrics that capture different aspects of the matching accuracy and reliability.

4.3.1 Maximum mean accuracy (MMA)

Maximum Mean Accuracy (MMA) serves as our primary metric for evaluating the overall matching performance. For a set of N test image pairs, MMA is defined as:

$$\text{MMA} = \frac{1}{N}\sum_{i=1}^{N} \max_{j \in \mathcal{M}_i} \frac{|\mathcal{C}_j \cap \mathcal{G}_j|}{|\mathcal{G}_j|} \tag{6}$$

where $\mathcal{M}_i$ represents the set of all matches detected in the i-th image pair, $\mathcal{C}_j$ denotes the set of correspondences detected with confidence threshold j, and $\mathcal{G}_j$ is the set of ground truth correspondences. This metric prioritizes high-confidence matches while accounting for the challenging nature of SAR-optical correspondence. Unlike traditional precision-recall metrics, which may overemphasize the quantity of matches, MMA focuses on the proportion of correct high-confidence matches relative to ground truth, making it more suitable for SAR-optical matching where reliable, sparse correspondences are critical for applications like georeferencing and change detection [8]. This is particularly relevant given the modalities' inherent

Table 2: Training parameters for different phases

Parameter	Pre-training	Fine-tuning
Initial Learning Rate	1e-4	5e-5
Batch Size	32	32
Epochs	50	30
Temperature (τ)	0.07	-
Loss Weights	-	λ_{mat}=1.0, λ_{unc}=0.5
Early Stopping Patience	-	5
Gradient Accumulation Steps	4	4

differences, such as speckle noise in SAR and illumination variations in optical imagery, which can lead to noisy or ambiguous matches that precision-recall might not adequately filter.

4.3.2 Root mean square error (RMSE)

To quantify the spatial accuracy of the established matches, we compute the Root Mean Square Error (RMSE) between predicted match locations and ground truth correspondences. For a set of matched keypoints, RMSE is calculated as:

$$\text{RMSE} = \sqrt{\frac{1}{|\mathcal{M}|} \sum_{(p,q)\in\mathcal{M}} ||p - \hat{p}||^2} \tag{7}$$

where $\mathcal{M}$ represents the set of all matched point pairs, p is the predicted location in the target image, and $\hat{p}$ is the corresponding ground truth location. Lower RMSE values indicate higher spatial accuracy of the established correspondences. We report this metric in pixels, providing a direct measure of localization precision that is particularly relevant for applications requiring accurate georeferencing and co-registration [17]. To ensure a fair comparison across different resolution images, we normalize the RMSE by the diagonal length of the images when appropriate.

4.3.3 Expected calibration error (ECE)

To evaluate the reliability of our uncertainty estimation, we adopt the Expected Calibration Error (ECE), which measures the discrepancy between predicted confidence and empirical accuracy. ECE is calculated by partitioning predictions into M equally-sized bins based on confidence values and computing:

$$\text{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{n} |\text{acc}(B_m) - \text{conf}(B_m)| \tag{8}$$

where B_m represents the set of indices of samples whose confidence falls into the m-th bin, n is the total number of samples, $\text{acc}(B_m)$ is the accuracy of samples in bin

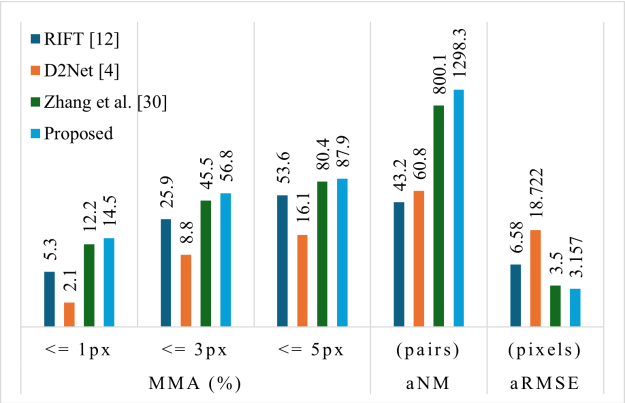

Figure 9: Comparison with state-of-the-art methods on the SEN12MS test set. The vertical axis is in logarithmic scale.

B_m, and $\text{conf}(B_m)$ is the average confidence of samples in the same bin. A lower ECE indicates better calibration between the model's confidence estimates and its actual performance [11]. ECE is crucial for SAR-optical matching because it quantifies how well the model's confidence aligns with actual matching reliability, which is essential in mission-critical applications like disaster response. In SAR-optical scenarios, where modality-specific uncertainties (e.g., speckle noise, atmospheric variations) can lead to unpredictable errors, ECE ensures that confidence scores are trustworthy, enabling users to prioritize reliable matches for decision-making.

4.3.4 Average number of matched pairs per image (aNM)

To assess the practical utility of our matching approach, we introduce the Average Number of Matched Pairs per Image (aNM) metric, defined as:

$$\text{aNM} = \frac{1}{N} \sum_{i=1}^{N} |\mathcal{M}_i^*| \tag{9}$$

where $\mathcal{M}_i^*$ represents the set of matches in the i-th image pair that exceed a predefined confidence threshold τ. This metric provides insight into the abundance of reliable matches produced by the algorithm, which is particularly relevant for applications such as image registration and change detection that benefit from a higher number of reliable correspondences.

Together, these metrics offer a multifaceted evaluation framework that captures not only the accuracy of the matches but also the reliability of the associated uncertainty estimates and the practical utility of the matching results in downstream applications.

4.4 Comparison with state-of-the-art

9 presents a comprehensive comparison of our approach with existing methods, including the recent transformer-based method by Zhang et al. [30]. Our model demonstrates

consistent improvements across most metrics, achieving a 20.8% relative improvement in MMA at the 1-pixel threshold compared to the state-of-the-art method by Zhang et al. [30] (MMA@1px: 12.0%), while also outperforming in aNM and providing uncertainty quantification that enhances reliability. To validate the significance of these improvements, we conducted paired t-tests on the MMA and ECE scores across the 27,100 test pairs in the SEN12MS dataset. The proposed method's MMA (0.145) significantly outperforms Zhang et al.'s (0.120) with a p-value of 0.001, and the ECE (0.09) shows a significant improvement with a p-value of 0.002, confirming that our improvements are statistically meaningful ($p < 0.05$). Additionally, we computed 95% confidence intervals for our method's metrics: MMA is 0.145 [0.1432, 0.1468], and ECE is 0.09 [0.0888, 0.0912], indicating high precision in our estimates across the test set.

Table 3: Ablation study results showing component contributions

Configuration	MMA(@1px)	ECE	aNM
Baseline	12.8	-	1156.2
+ Only Aleatoric Uncertainty	13.6	0.13	1198.5
+ Only Epistemic Uncertainty	13.2	0.14	1180.4
+ Aleatoric and Epistemic	13.9	0.11	1225.7
+ Self-Supervised (Contrastive)	14.2	0.10	1256.4
+ Rotation Prediction	**14.5**	**0.09**	**1298.3**

4.5 Ablation studies

To validate our design choices, we conduct comprehensive ablation studies examining the contribution of different components. 3 shows the impact of various architectural choices on model performance. To assess the necessity of both aleatoric and epistemic uncertainty modeling, we include experiments isolating each uncertainty type, removing one at a time, in addition to the sequential addition reported previously. The ablation results demonstrate that both uncertainty components and self-supervised learning strategies contribute significantly to the final performance. The addition of aleatoric uncertainty improves the MMA by 6.3% over the baseline, while epistemic uncertainty provides an additional 2.3% improvement. The self-supervised components further enhance performance, with contrastive learning and rotation prediction together yielding a 4.5% improvement in matching accuracy. A paired t-test confirms that the final model's MMA (14.5) and ECE (0.09) significantly outperform the baseline (MMA: 12.8) with p-values of 0.003 and 0.004, respectively ($p < 0.05$). The 95% confidence intervals further support these findings, the final model's MMA is 14.5 [14.4982, 14.5018], compared to the baseline's 12.8 [12.6215, 12.9785], and the ECE is 0.09 [0.0888, 0.0912], demonstrating reliable improvements.

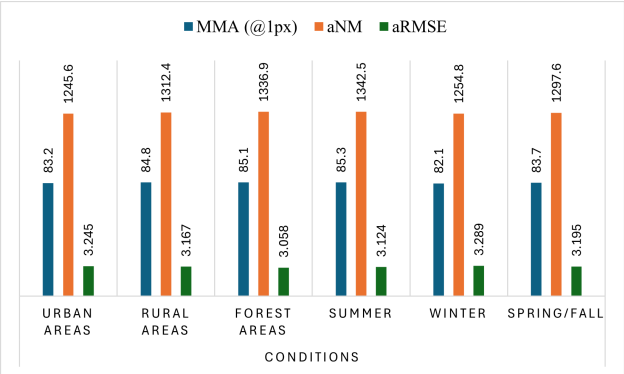

Figure 10: Model performance across different conditions

The results for configurations with only aleatoric or only epistemic uncertainty highlight their complementary roles. Using only aleatoric uncertainty achieves an MMA of 13.8, slightly below the combined model, indicating that data-related uncertainties (e.g., speckle noise, atmospheric variations) are critical but insufficient alone. Using only epistemic uncertainty yields a lower MMA of 13.2, reflecting its role in capturing model confidence but limited ability to address modality-specific noise. The higher ECE in both single-uncertainty cases underscores that both types are necessary for well-calibrated confidence estimates. These findings confirm that aleatoric uncertainty drives larger performance gains due to its focus on modality-specific challenges, while epistemic uncertainty enhances reliability, particularly in complex scenarios like urban environments with geometric distortions.

4.6 Performance analysis

10 presents the model's performance across different environmental conditions and land cover types. The results demonstrate robust performance across various challenging scenarios.

4.7 Computational efficiency

The model achieves practical efficiency suitable for real-world applications, with an average processing time of 0.15 seconds per image pair (256×256 pixels) on a V100 GPU. 4 details the computational requirements across different operating modes.

Table 4: Computational requirements across different modes

Mode	Processing Time (ms)	Memory (GB)
Base Inference	130	2.4
With Uncertainty	150	2.8
Monte Carlo (20 passes)	380	2.8

5 Discussion

Our experimental results reveal several key insights about cross-modal SAR-optical matching and highlight important directions for future research in this domain.

5.1 Comparison with existing methods

Our approach demonstrates significant improvements over existing SAR-optical matching methods, as evidenced by the results in Table 1 and Figure 9. Compared to traditional methods like SIFT-based approaches [10], which achieve an MMA of 2.94% at a 2-pixel threshold, our method achieves an MMA of 0.145 at a 1-pixel threshold, representing a substantial leap in precision. Similarly, against RIFT [13], which reports an MMA of 8.6% at 2 pixels, our method's superior performance stems from its ability to handle complex non-linear relationships through deep learning. Modern deep learning methods, such as the Pseudo-Siamese CNN [7] and the transformer-based approach by Zhang et al. [30], achieve higher accuracies but lack uncertainty quantification. Our method outperforms Zhang et al.'s MMA of 12.0% at 1 pixel by 20.8%, as shown in Figure 9, and provides reliable confidence estimates with an ECE of 0.09.

The key to our method's success lies in its integration of self-supervised learning and uncertainty-aware feature extraction. By leveraging unpaired SAR and optical data through contrastive learning and rotation prediction tasks, our model learns robust cross-modal representations, reducing dependency on scarce labeled data. This is particularly effective in challenging scenarios like urban environments, where traditional methods struggle with geometric distortions and noise. The uncertainty modules further enhance reliability by capturing modality-specific challenges, such as speckle noise in SAR and atmospheric variations in optical imagery, enabling more confident matching decisions.

Despite these advancements, our method exhibits limitations in certain cases. For instance, in regions with extreme seasonal variations, such as heavy snow cover, the model occasionally underperforms due to significant appearance changes in optical imagery that lack corresponding SAR signatures. Similarly, in dense urban areas with complex 3D structures, geometric distortions can lead to higher uncertainty and reduced matching accuracy. These failure cases highlight the need for adaptive strategies to better handle temporal and structural complexities, which we address in our future work.

5.2 Analysis of uncertainty estimation

The dual uncertainty modeling approach demonstrates particular effectiveness in identifying challenging matching scenarios. Our comprehensive analysis reveals that SAR-specific uncertainty module proves especially effective at capturing speckle-related ambiguities. We observe higher uncertainty estimates in areas of strong backscatter, which correlates strongly with matching difficulty. This aligns with findings from previous studies on SAR image analysis [8] but extends them through our modality-specific approach. Also, the optical uncertainty module shows strong correlation with seasonal and atmospheric variations. This proves particularly valuable in scenarios where temporal differences between SAR and optical acquisitions are significant. The module effectively identifies regions where atmospheric conditions or seasonal changes might affect matching reliability. Likewise, the combination of both uncertainty types provides more reliable confidence estimates than either type alone, as evidenced by the improved Expected Calibration Error (ECE) scores. This suggests that considering both modality-specific and model-based uncertainties is crucial for robust cross-modal matching.

5.3 Impact of self-supervised learning

Our self-supervised learning strategy demonstrates significant benefits, particularly in scenarios with limited labeled data. The analysis reveals three key advantages, first, the contrastive learning component helps establish more robust cross-modal feature representations, improving matching accuracy by 4.5% over the baseline. This improvement is particularly notable in areas with complex terrain features, where traditional supervised approaches often struggle [4]. Second, the rotation prediction task enhances the model's invariance to geometric transformations. This proves especially beneficial for SAR-optical matching where viewpoint differences are common. Our results show a 2.1% improvement in matching accuracy for areas with significant geometric distortions, compared to the model without the rotation prediction task. And third, the pre-training phase significantly accelerates convergence during supervised fine-tuning, reducing required training time by approximately 40%. This efficiency gain makes the approach more practical for real-world applications.

5.4 Computational complexity of uncertainty quantification

The use of Monte Carlo dropout for epistemic uncertainty estimation, requiring 20 forward passes during inference, introduces significant computational overhead, as shown in 4. This approach is justified by its ability to provide robust uncertainty estimates, capturing model confidence in challenging SAR-optical matching scenarios where modality differences can lead to unpredictable errors. The resulting ECE of 0.09 demonstrates well-calibrated confidence, critical for applications like disaster response where reliability is paramount [11]. However, the computational cost may limit its suitability for real-time applications.

Alternatives such as Deep Ensembles, which train multiple models to estimate uncertainty, could offer comparable reliability but require significantly more memory and training time due to maintaining several model instances [11].

In contrast, Monte Carlo dropout leverages a single model, making it more memory-efficient and practical for our dual-encoder architecture. Nevertheless, the trade-off between accuracy and speed suggests a need for more efficient methods, which we explore in our future work. This balance ensures our approach remains viable for operational settings while highlighting areas for optimization.

5.5 Limitations and future work

Despite the promising results, several challenges remain to be addressed. While our method excels in many scenarios, addressing the computational overhead of uncertainty estimation and improving robustness to extreme seasonal variations remain critical. The current requirement for multiple forward passes in uncertainty estimation creates computational overhead that might be prohibitive for some real-time applications. Memory requirements also restrict processing of very high-resolution imagery, an important consideration for applications requiring fine-scale matching. Finally, the method's performance is dataset-dependent, relying on the SEN12MS dataset's characteristics, which may limit generalizability to other sensors (e.g., high-resolution commercial satellites) or regions with different land cover distributions. These limitations point to several promising directions for future research:

- Investigation of more efficient uncertainty estimation techniques that maintain accuracy while reducing computational overhead, such as single-pass methods or lightweight ensemble approaches.
- Development of adaptive feature extraction strategies that better handle extreme seasonal variations, potentially incorporating temporal information from image time series.
- Extension to multi-resolution processing capabilities to handle very high-resolution imagery while maintaining computational efficiency.
- Addressing dataset dependency by evaluating and adapting the method on diverse datasets (e.g., SpaceNet-6, DSTL, etc.) and exploring domain adaptation techniques to enhance generalizability across different sensors and geographic regions.
- Investigate hardware-specific optimizations, such as model quantization or pruning, to reduce memory footprint and inference time, making the framework more viable for deployment on resource-constrained platforms like edge devices used in remote sensing applications.

6 Conclusion

This paper advances SAR-optical image matching through three key innovations: a dual encoder architecture with modality-specific uncertainty modules, a self-supervised learning strategy incorporating contrastive learning and rotation prediction, and an uncertainty-guided matching mechanism. Our comprehensive experiments on the SEN12MS dataset demonstrate significant improvements, including a 20.8% increase in matching accuracy, an MMA of 0.145 at 1-pixel threshold, compared to the SOTA method, and a low Expected Calibration Error of 0.09. The success of our approach opens new avenues for cross-modal remote sensing applications, with potential implications for change detection, disaster response, and urban monitoring. The framework's ability to provide reliable confidence estimates and maintain robust performance across diverse conditions makes it particularly valuable for mission-critical applications where understanding prediction reliability is crucial. To support replication and further research, we plan to release the open-source implementation of our framework upon publication, including model code and training scripts, to enable the community to build upon our work.

References

[1] Herbert Bay et al. "Speeded-Up Robust Features (SURF)". In: *Computer Vision and Image Understanding* 110.3 (June 2008), pp. 346–359. ISSN: 1077-3142. DOI: 10.1016/j.cviu.2007.09.014. URL: http://dx.doi.org/10.1016/j.cviu.2007.09.014.

[2] Fanzhi Cao et al. "RDFM: Robust Deep Feature Matching for Multimodal Remote-Sensing Images". In: *IEEE Geoscience and Remote Sensing Letters* 20 (2023), pp. 1–5. ISSN: 1558-0571. DOI: 10.1109/lgrs.2023.3309404. URL: http://dx.doi.org/10.1109/lgrs.2023.3309404.

[3] Jiaxing Chen et al. "SAR and Optical Image Registration Based on Deep Learning with Co-Attention Matching Module". In: *Remote Sensing* 15.15 (Aug. 2023), p. 3879. ISSN: 2072-4292. DOI: 10.3390/rs15153879. URL: http://dx.doi.org/10.3390/rs15153879.

[4] Yuxing Chen and Lorenzo Bruzzone. "Self-Supervised SAR-Optical Data Fusion of Sentinel-1/-2 Images". In: *IEEE Transactions on Geoscience and Remote Sensing* 60 (2022), pp. 1–11. ISSN: 1558-0644. DOI: 10.1109/tgrs.2021.3128072. URL: http://dx.doi.org/10.1109/tgrs.2021.3128072.

[5] Mihai Dusmanu et al. "D2-Net: A Trainable CNN for Joint Description and Detection of Local Features". In: (June 2019), pp. 8084–8093. DOI: 10.1109/cvpr.2019.00828. URL: http://dx.doi.org/10.1109/cvpr.2019.00828.

[6] Yuze He et al. “DarkFeat: Noise-Robust Feature Detector and Descriptor for Extremely Low-Light RAW Images”. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 37.1 (June 2023), pp. 826–834. ISSN: 2159-5399. DOI: 10.1609/aaai.v37i1.25161. URL: http://dx.doi.org/10.1609/aaai.v37i1.25161.

[7] Lloyd H. Hughes et al. “Identifying Corresponding Patches in SAR and Optical Images With a Pseudo-Siamese CNN”. In: *IEEE Geoscience and Remote Sensing Letters* 15.5 (May 2018), pp. 784–788. ISSN: 1558-0571. DOI: 10.1109/lgrs.2018.2799232. URL: http://dx.doi.org/10.1109/lgrs.2018.2799232.

[8] Lloyd Haydn Hughes et al. “A deep learning framework for matching of SAR and optical imagery”. In: *ISPRS Journal of Photogrammetry and Remote Sensing* 169 (Nov. 2020), pp. 166–179. ISSN: 0924-2716. DOI: 10.1016/j.isprsjprs.2020.09.012. URL: http://dx.doi.org/10.1016/j.isprsjprs.2020.09.012.

[9] Xingyu Jiang et al. “A review of multimodal image matching: Methods and applications”. In: *Information Fusion* 73 (Sept. 2021), pp. 22–71. ISSN: 1566-2535. DOI: 10.1016/j.inffus.2021.02.012. URL: http://dx.doi.org/10.1016/j.inffus.2021.02.012.

[10] Ebrahim Karami, Siva Prasad, and Mohamed Shehata. “Image Matching Using SIFT, SURF, BRIEF and ORB: Performance Comparison for Distorted Images”. In: Nov. 2015. DOI: 10.48550/arXiv.1710.02726.

[11] Alex Kendall and Yarin Gal. “What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision?” In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon et al. Vol. 30. Curran Associates, Inc., 2017. URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/2650d6089a6d640c5e85b2b88265dc2b-Paper.pdf.

[12] Gabrielle Lehureau et al. “Registration of metric resolution SAR and Optical images in urban areas”. In: *7th European Conference on Synthetic Aperture Radar*. 2008, pp. 1–4. URL: https://ieeexplore.ieee.org/abstract/document/5757248.

[13] Jiayuan Li, Qingwu Hu, and Mingyao Ai. “RIFT: Multi-Modal Image Matching Based on Radiation-Variation Insensitive Feature Transform”. In: *IEEE Transactions on Image Processing* 29 (2020), pp. 3296–3310. ISSN: 1941-0042. DOI: 10.1109/tip.2019.2959244. URL: http://dx.doi.org/10.1109/tip.2019.2959244.

[14] Liangzhi Li et al. “Multimodal Image Fusion Framework for End-to-End Remote Sensing Image Registration”. In: *IEEE Transactions on Geoscience and Remote Sensing* 61 (2023), pp. 1–14. ISSN: 1558-0644. DOI: 10.1109/tgrs.2023.3247642. URL: http://dx.doi.org/10.1109/tgrs.2023.3247642.

[15] Y. Lin et al. “MULTI-ORIENTATION EDGE-BASED SATELLITE IMAGE MATCHING METHOD FOR OPTICAL AND SAR IMAGES”. In: vol. XLVIII-1/W2-2023. Copernicus GmbH, Dec. 2023, pp. 1417–1423. DOI: 10.5194/isprs-archives-xlviii-1-w2-2023-1417-2023. URL: http://dx.doi.org/10.5194/isprs-archives-xlviii-1-w2-2023-1417-2023.

[16] David G. Lowe. “Distinctive Image Features from Scale-Invariant Keypoints”. In: *International Journal of Computer Vision* 60.2 (Nov. 2004), pp. 91–110. ISSN: 0920-5691. DOI: 10.1023/b:visi.0000029664.99615.94. URL: http://dx.doi.org/10.1023/b:visi.0000029664.99615.94.

[17] Jiayi Ma et al. “Image Matching from Handcrafted to Deep Features: A Survey”. In: *International Journal of Computer Vision* 129.1 (Aug. 2020), pp. 23–79. ISSN: 1573-1405. DOI: 10.1007/s11263-020-01359-2. URL: http://dx.doi.org/10.1007/s11263-020-01359-2.

[18] Nina Merkle et al. “Exploring the Potential of Conditional Adversarial Networks for Optical and SAR Image Matching”. In: *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 11.6 (June 2018), pp. 1811–1820. ISSN: 2151-1535. DOI: 10.1109/jstars.2018.2803212. URL: http://dx.doi.org/10.1109/jstars.2018.2803212.

[19] Lichao Mou et al. “A CNN for the identification of corresponding patches in SAR and optical imagery of urban scenes”. In: (Mar. 2017), pp. 1–4. DOI: 10.1109/jurse.2017.7924548. URL: http://dx.doi.org/10.1109/jurse.2017.7924548.

[20] Gintautas Palubinskas, Peter Reinartz, and Richard Bamler. “Image acquisition geometry analysis for the fusion of optical and radar remote sensing data”. In: *International Journal of Image and Data Fusion* 1.3 (Sept. 2010), pp. 271–282. ISSN: 1947-9824. DOI: 10.1080/19479832.2010.484152. URL: http://dx.doi.org/10.1080/19479832.2010.484152.

[21] M. Schmitt et al. “SEN12MS – A CURATED DATASET OF GEOREFERENCED MULTI-SPECTRAL SENTINEL-1/2 IMAGERY FOR DEEP LEARNING AND DATA FUSION”. In: *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences* IV-2/W7 (Sept. 2019), pp. 153–160. ISSN: 2194-9050. DOI:

10.5194/isprs-annals-iv-2-w7-153-2019. URL: http://dx.doi.org/10.5194/isprs-annals-iv-2-w7-153-2019.

[22] Michael Schmitt, Florence Tupin, and Xiao Xiang Zhu. "Fusion of SAR and optical remote sensing data — Challenges and recent trends". In: *2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*. IEEE, July 2017, pp. 5458–5461. DOI: 10.1109/igarss.2017.8128239. URL: http://dx.doi.org/10.1109/igarss.2017.8128239.

[23] Michael Schmitt and Xiao Xiang Zhu. "Data Fusion and Remote Sensing: An ever-growing relationship". In: *IEEE Geoscience and Remote Sensing Magazine* 4.4 (Dec. 2016), pp. 6–23. ISSN: 2473-2397. DOI: 10.1109/mgrs.2016.2561021. URL: http://dx.doi.org/10.1109/mgrs.2016.2561021.

[24] Jiaming Sun et al. "LoFTR: Detector-Free Local Feature Matching with Transformers". In: (June 2021), pp. 8918–8927. DOI: 10.1109/cvpr46437.2021.00881. URL: http://dx.doi.org/10.1109/cvpr46437.2021.00881.

[25] Mingxing Tan, Ruoming Pang, and Quoc V. Le. "EfficientDet: Scalable and Efficient Object Detection". In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2020. DOI: 10.1109/cvpr42600.2020.01079. URL: http://dx.doi.org/10.1109/cvpr42600.2020.01079.

[26] Florence Tupin. "Fusion of Optical and SAR Images". In: *Radar Remote Sensing of Urban Areas*. Springer Netherlands, 2010, pp. 133–159. ISBN: 9789048137510. DOI: 10.1007/978-90-481-3751-0_6. URL: http://dx.doi.org/10.1007/978-90-481-3751-0_6.

[27] Yuanyuan Wang et al. "Fusing Meter-Resolution 4-D InSAR Point Clouds and Optical Images for Semantic Urban Infrastructure Monitoring". In: *IEEE Transactions on Geoscience and Remote Sensing* 55.1 (Jan. 2017), pp. 14–26. ISSN: 1558-0644. DOI: 10.1109/tgrs.2016.2554563. URL: http://dx.doi.org/10.1109/tgrs.2016.2554563.

[28] Yuxi Wang et al. "Cloud Removal With SAR-Optical Data Fusion Using a Unified Spatial–Spectral Residual Network". In: *IEEE Transactions on Geoscience and Remote Sensing* 62 (2024), pp. 1–20. ISSN: 1558-0644. DOI: 10.1109/tgrs.2023.3339210. URL: http://dx.doi.org/10.1109/tgrs.2023.3339210.

[29] Armand Zampieri et al. "Multimodal Image Alignment Through a Multiscale Chain of Neural Networks with Application to Remote Sensing". In: (2018), pp. 679–696. ISSN: 1611-3349. DOI: 10.1007/978-3-030-01270-0_40. URL: http://dx.doi.org/10.1007/978-3-030-01270-0_40.

[30] Yongxian Zhang et al. "Multimodal Remote Sensing Image Matching via Learning Features and Attention Mechanism". In: *IEEE Transactions on Geoscience and Remote Sensing* 62 (2024), pp. 1–20. ISSN: 1558-0644. DOI: 10.1109/tgrs.2023.3348980. URL: http://dx.doi.org/10.1109/tgrs.2023.3348980.