O2SMatch: Cross-modality Knowledge Transferring Induced Optical and SAR Image Matching

Jiahang Song and Ganggang Dong

***Abstract*—Accurate matching between SAR and optical images remains challenging due to the significant discrepancy in imaging mechanisms and visual characteristics. To address this issue, this paper proposes a cross-modality knowledge transferring framework for SAR–optical image matching. First, an adversarially accelerated diffusion model is introduced to construct an intermediate domain between SAR and optical images, effectively bridging the modality gap. Based on the translated representations, a cascaded cross-sensor matching architecture is developed using a densely connected convolutional network equipped with hierarchical attention mechanisms, including cross-channel, cross-tensor, and cross-kernel attention. Extensive experiments on measured datasets demonstrate that the proposed method significantly outperforms competitive approaches.**

I. Introduction

Radar and optical sensors provide complementary information for Earth observation. SAR imaging is capable of acquiring high-resolution images under all-weather and day-and-night conditions [1], while optical images offer rich spectral and semantic information. Accurate SAR–optical image matching is therefore a fundamental prerequisite for many downstream remote sensing applications. However, significant differences in imaging mechanisms and radiometric characteristics make cross-sensor matching highly challenging [2]. Existing solutions can be broadly categorized into model-driven and learning-driven methods.

A. Model-driven methods

Early image matching methods mainly relied on hand-crafted features derived from empirical models. The most representative approach is the scale-invariant feature transform (SIFT) [3], which detects keypoints in scale space and describes local regions using gradient orientation histograms. Based on SIFT, numerous variants have been proposed. Brown and Lowe applied SIFT to panoramic image stitching [4], while Bay et al. introduced SURF as a computationally efficient alternative [5]. Rublee et al. proposed ORB, a fast binary descriptor suitable for real-time matching [6]. Liu et al. further introduced a log-polar sampling strategy to improve scale and rotation invariance [7].

Although satisfactory performance can be achieved in intra-modal scenarios, the generalization ability of hand-crafted features across heterogeneous sensors is limited. As illustrated in Fig. 1, severe misalignment often occurs when applying conventional descriptors to SAR–optical matching. This is mainly because gradient-based descriptors are sensitive to speckle noise and radiometric distortions. The limitations of model-driven methods can be summarized as follows:

- sensitivity to modality-specific artifacts;
- poor robustness to large appearance variations;
- lack of shared representation modeling across sensors.

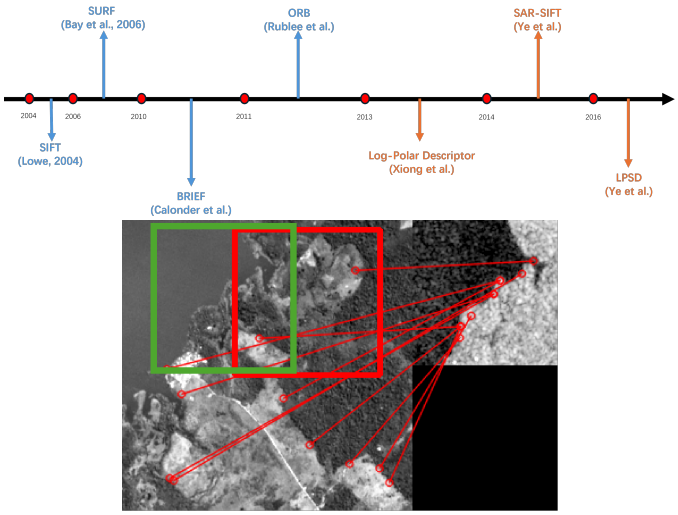

Fig. 1: The demonstration of SIFT for optical and SAR image matching. The severe misalignment was produced.

B. Learning-driven methods

With the rapid development of deep learning, recent studies have shifted toward learning robust feature representations directly from data. DeTone et al. proposed SuperPoint to jointly learn interest point detection and descriptor extraction in a self-supervised manner [8]. Mishchuk et al. introduced HardNet, which improves descriptor discrimination through hard negative mining [9]. Building upon these works, various deep architectures have been developed for image matching, as summarized in Fig. 2.

Recent approaches further incorporate transformer-based architectures to enhance global context modeling. Sun et al. proposed a detector-free dense matching framework using self- and cross-attention [10]. Lindenberger et al. introduced a lightweight transformer for efficient sparse matching [11], while Wang et al. developed MatchFormer by integrating transformers into the matching pipeline [12]. Despite their success in optical image matching, learning-driven methods still exhibit limited generalization in SAR–optical scenarios. Most

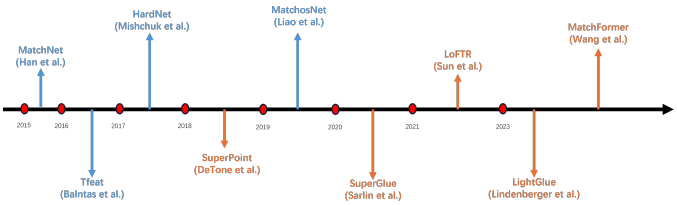

Fig. 2: The representative learning-driven image matching methods.

existing approaches emphasize matching architecture design while neglecting the intrinsic spectral and physical differences between sensors. As a result, learned descriptors often fail to capture SAR-specific scattering patterns, leading to sparse or erroneous correspondences in complex scenes. Recent IJCNN work has demonstrated the effectiveness of data-driven models for SAR image processing. FocusNet [13] focuses on SAR autofocus via deep learning, while IFENet [14] addresses SAR image formation and enhancement under missing-echo conditions. These studies highlight the importance of learning-based modeling for SAR-specific degradations and motivate modality-aware designs for cross-sensor matching.

C. Our Solution

To address the aforementioned challenges, this paper proposes a cross-modality knowledge transferring induced image matching framework. Unlike existing approaches, the SAR–optical matching task is decoupled into two sequential stages: modality translation and cascaded matching.

- *Modality translation:* An adversarially accelerated diffusion model is introduced to construct an intermediate domain between SAR and optical images, effectively bridging the modality gap.
- *Cascaded matching:* A densely connected convolutional network equipped with hierarchical attention mechanisms is developed for cross-sensor matching. Cross-channel, cross-tensor, and cross-kernel attention are jointly integrated to enhance semantic feature extraction.

The overall framework is illustrated in Fig. 3. The major contributions of this paper are summarized as follows:

- A two-stage framework is proposed to decouple SAR–optical matching into modality translation and cascaded matching.
- An adversarially accelerated diffusion model is introduced to bridge the large modality gap between SAR and optical images.
- A hierarchical attention based densely connected network is developed for robust cross-sensor image matching.

II. The Image Matching

Image matching aims to establish reliable correspondences between images acquired under different viewpoints, times, or sensing modalities. A typical pipeline consists of feature detection, descriptor extraction, feature matching, and geometric verification. Classical methods such as SIFT employ hand-crafted descriptors with scale and rotation invariance, followed by nearest-neighbor matching and RANSAC-based outlier rejection.

While these approaches are effective for intra-modal matching, they often fail under SAR–optical settings due to severe radiometric distortion and speckle noise.

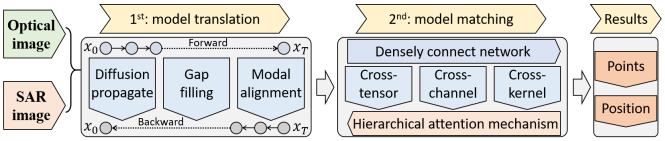

Fig. 3: The overall diagram of proposed method.

III. The Proposed Method

To mitigate the large appearance discrepancy between SAR and optical imagery, the cross-sensor image matching task is decoupled into two sequential sub-tasks: cross-modality translation and cascaded matching. The former aims to construct an intermediate modality that bridges SAR and optical domains, while the latter focuses on accurate correspondence estimation via hierarchical attention learning. The overall framework is illustrated in Fig. 4.

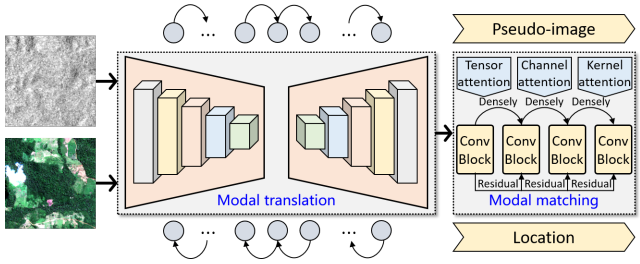

Fig. 4: The learning framework of the proposed method.

A. Cross-Modality Translation

The first stage targets the translation from speckle-contaminated SAR images to appearance-consistent pseudo-optical images. We adopt a conditional latent diffusion model enhanced by adversarial learning, referred to as DiffOS, to construct an intermediate modality. SAR images serve as structural guidance throughout the diffusion process. The overall architecture is shown in Fig. 5.

We adopt a conditional latent diffusion model to progressively translate SAR images into pseudo-optical representations via a noise-to-data denoising process. The principle is simply shown in Fig. 6.

1) Forward process: Given a real optical sample $\mathbf{x}_0 \sim q(\mathbf{x})$, the forward diffusion process progressively injects Gaussian noise,

$$q(\mathbf{x}_t \mid \mathbf{x}_{t-1}) = \mathcal{N}\left(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t \mathbf{I}\right), \quad (1)$$

which admits the closed-form expression

$$q(\mathbf{x}_t \mid \mathbf{x}_0) = \mathcal{N}\left(\mathbf{x}_t; \sqrt{\bar{\alpha}_t}\mathbf{x}_0, (1-\bar{\alpha}_t)\mathbf{I}\right), \quad (2)$$

where $\bar{\alpha}_t = \prod_{i=1}^{t}(1-\beta_i)$.

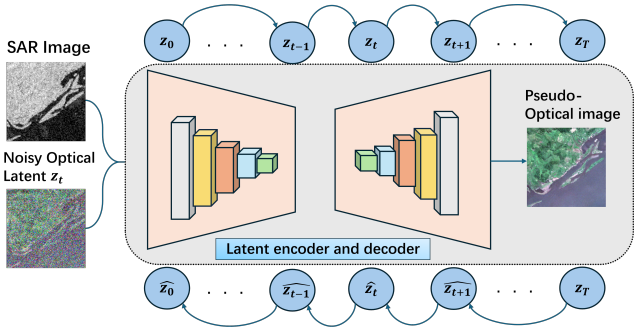

Fig. 5: Architecture of the proposed DiffOS framework for SAR-to-optical translation.

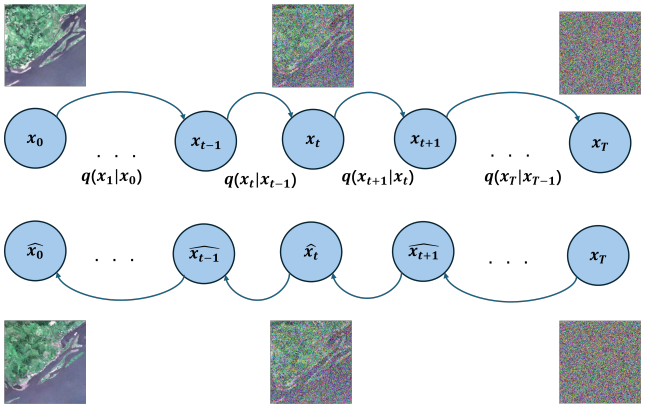

Fig. 6: Illustration of the forward and reverse process of base diffusion model.

2) Reverse process with SAR guidance: To enable cross-modality translation, the reverse denoising process is conditioned on the SAR images,

$$p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{s}) = \mathcal{N}(\mathbf{x}_{t-1}; \boldsymbol{\mu}_\theta(\mathbf{x}_t, \mathbf{s}, t), \boldsymbol{\Sigma}_t)\,. \tag{3}$$

The model predicts the noise term $\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, \mathbf{s}, t)$, while operating in a latent space for computational efficiency. At each step, the noisy optical latent $\mathbf{z}_t$ is concatenated with the normalized SAR image,

$$\mathbf{h}_t = \oplus(\mathbf{z}_t, \mathbf{s}), \tag{4}$$

ensuring SAR-derived geometric structures guide the denoising trajectory. The final latent $\mathbf{z}_0$ is decoded to obtain the pseudo-optical image $\hat{\mathbf{x}}_0$.

3) Adversarial enhancement: To further improve perceptual realism, an image-domain adversarial loss is introduced,

$$\mathcal{L}_{\text{adv}} = \mathbb{E}_{\mathbf{x}_0}[\log \mathcal{D}(\mathbf{x}_0)] + \mathbb{E}_{\mathbf{s}}[\log(1 - \mathcal{D}(\hat{\mathbf{x}}_0))], \tag{5}$$

which encourages the translated images to follow realistic optical distributions while preserving SAR-induced structural constraints.

B. Cross-Modality Matching

The generated pseudo-optical images facilitate the subsequent matching stage. A dedicated network, termed *CascadeMatch*, is designed to learn discriminative cross-modality descriptors.

1) Architecture: CascadeMatch consists of three densely connected convolutional blocks followed by a prediction module. Each block integrates hierarchical attention mechanisms, including cross-channel attention (CCA), cross-kernel attention (CKA), and cross-tensor attention (CTA), to jointly address channel imbalance, scale inconsistency, and cross-modality geometric variation. The overall architecture is shown in Fig. 7.

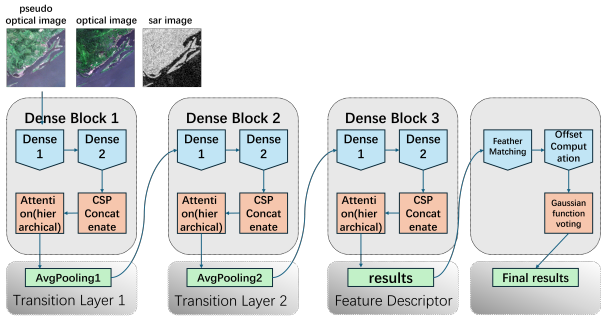

Fig. 7: Overall architecture of CascadeMatch.

2) Hierarchical attention: CCA recalibrates channel-wise responses to suppress modality-specific noise. CKA adaptively aggregates multi-scale convolutional features to handle spatial-scale variations. CTA models cross-descriptor interactions to enforce geometric consistency across modalities.

3) Prediction: Given descriptor sets $\mathcal{D}^A = \{\mathbf{d}_i^A\}$ and $\mathcal{D}^B = \{\mathbf{d}_j^B\}$, a similarity matrix is computed as

$$S_{ij} = -\|\mathbf{d}_i^A - \mathbf{d}_j^B\|_2. \tag{6}$$

For each descriptor, a k-NN search is performed and weighted voting is applied to aggregate spatial offsets,

$$V(\Delta\mathbf{p}) = \sum_i \sum_{j \in \mathcal{N}_k(i)} \alpha_{ij} \exp\left(-\frac{\|\Delta\mathbf{p} - \Delta\mathbf{p}_{ij}\|_2^2}{2\sigma^2}\right). \tag{7}$$

The dominant peak yields the final correspondence, enforcing global geometric consistency.

C. Learning Strategy

The network is trained using a joint global and local objective. A margin-based global loss enforces descriptor separation,

$$\mathcal{L}_{\text{g}} = \frac{1}{N} \sum_i \max\big(0,\, \delta + d(\mathbf{f}_i^o, \mathbf{f}_i^s) - d(\mathbf{f}_i^o, \mathbf{f}_i^n)\big), \tag{8}$$

while a cosine-based local loss improves fine-grained discrimination,

$$\mathcal{L}_{\text{l}} = -\frac{1}{N} \sum_i \log \frac{\exp(\gamma \cos(\theta_i + \mu))}{\exp(\gamma \cos(\theta_i + \mu)) + \sum_{k \in \mathcal{N}_i} \exp(\gamma \cos(\theta_{i,k}))}. \tag{9}$$

The final objective is

$$\mathcal{L}_{\text{DiffOS}} = \mathcal{L}_{\text{g}} + \lambda \mathcal{L}_{\text{l}}. \tag{10}$$

IV. EXPERIMENTS AND ANALYSIS

This section validates the effectiveness of the proposed DiffOS framework through extensive experiments. We first introduce the experimental settings and evaluation metrics. Then, the cross-modality translation module is quantitatively analyzed, followed by comprehensive ablation studies and comparisons with state-of-the-art matching methods.

A. Experimental Setting

1) Dataset: The proposed method is verified on two publicly released benchmark datasets, SEN1-2 and OSDataset. **SEN1-2.** SEN1-2 [15] contains co-registered SAR images from Sentinel-1 and optical images from Sentinel-2, with a spatial resolution of 256×256 pixels. Both summer and winter subsets are used to evaluate robustness under seasonal variations. Each subset contains approximately 1000 paired samples covering diverse land-cover types. The data are split into training, validation, and test sets following a consistent protocol. **OSDataset.** OSDataset [16], [17] provides SAR images acquired by GF-3 in spotlight mode and corresponding optical images from Google Earth. Compared with SEN1-2, OSDataset exhibits weaker optical texture, stronger noise, and larger modality discrepancies, making it more challenging for cross-modal matching. For matching evaluation, a 128×128 patch is randomly cropped from either the pseudo-optical image (forward matching) or the real optical image (reverse matching) and aligned with the corresponding 256×256 reference image. The known crop offset serves as ground truth.

2) Evaluation Metrics: Since DiffOS decouples cross-sensor matching into translation and matching stages, both are evaluated separately. **Modality translation**. Translation quality is assessed using PSNR, SSIM, and MSE between pseudo-optical and real optical images. **Image matching**. Matching accuracy is measured by the mean Intersection over Union (mIoU) between predicted and ground-truth regions. We also report the proportion of samples achieving IoU > 0.9 to reflect high-precision localization.

B. The verification of cross-modality translation

We first evaluate the proposed diffusion-based translation module on SEN1-2 and OSDataset. Quantitative results are summarized in Tables I–III.

TABLE I: The knowledge transferring results on SEN1-2 (Summer).

(a) The comparison with state-of-the-art

Method	PSNR (dB)	SSIM	MSE
Pix2Pix	12.21	0.1210	4485.36
CycleGAN	12.79	0.1228	4149.43
DiffOS	**13.18**	**0.1714**	**3520.85**

(b) The bidirectional generation experiments

Metric	SAR–Opt	Pseudo–Opt (DiffOS)
Mean PSNR	6.55	**13.18**
Mean SSIM	0.0621	**0.1714**
Mean MSE	14940.26	**3520.85**

TABLE II: The knowledge transferring results on SEN1-2 (Winter).

(a) The comparison with state-of-the-art

Method	PSNR (dB)	SSIM	MSE
Pix2Pix	14.20	0.1927	2969.59
CycleGAN	12.03	0.1006	4893.57
DiffOS	**14.90**	**0.2823**	**2350.96**

(b) The bidirectional generation experiments

Metric	SAR–Opt	Pseudo–Opt (DiffOS)
Mean PSNR	9.48	**14.90**
Mean SSIM	-0.0377	**0.2823**
Mean MSE	7840.51	**2350.96**

Compared with direct SAR–optical comparison, the translated pseudo-optical images significantly improve all pixel-level similarity metrics, including higher PSNR and SSIM and substantially lower MSE, demonstrating that the proposed diffusion-based translation effectively reduces cross-modal appearance discrepancies. On the SEN1-2 summer subset, DiffOS improves PSNR by 6.63 dB and SSIM by 0.109, while reducing MSE by 76.4% relative to direct SAR–optical comparison. All translation-based methods outperform direct comparison, confirming the necessity of introducing an intermediate modality, while DiffOS consistently achieves the best performance among all methods.In addition to pixel-level metrics, we further visualize the cross-modal feature distributions using t-SNE to provide an intuitive understanding of the modality gap reduction. As shown in Fig. 8, features extracted from SAR and optical images form two clearly separated clusters, indicating a severe discrepancy in the original feature space. After diffusion-based translation, the pseudo-optical features are significantly closer to the optical cluster and partially overlap with it, while remaining well separated from the SAR features. This observation suggests

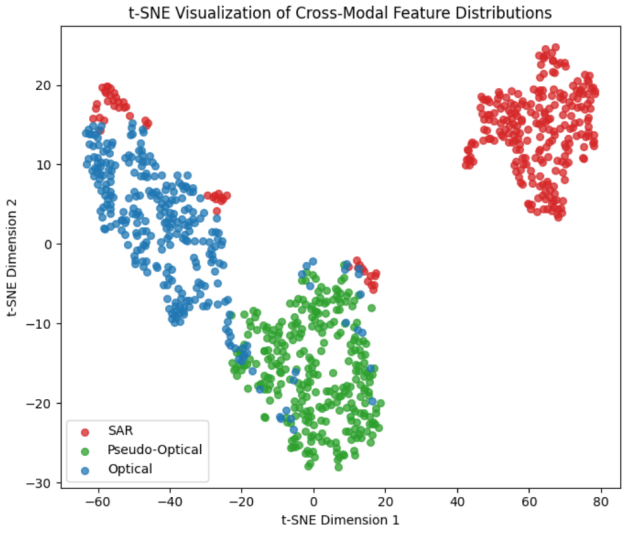

Fig. 8: t-SNE visualization of cross-modal feature distributions

that DiffOS effectively maps SAR representations into an

TABLE III: The knowledge transferring results on OSDataset.

(a) The comparison with state-of-the-art

Method	PSNR (dB)	SSIM	MSE
Pix2Pix	16.69	0.2390	1562.55
CycleGAN	17.39	0.2901	1331.41
DiffOS	**17.53**	**0.3237**	**1266.62**

(b) The bidirectional generation experiments

Metric	SAR–Opt	Pseudo–Opt (DiffOS)
Mean PSNR	11.51	**17.53**
Mean SSIM	0.0606	**0.3237**
Mean MSE	5186.07	**1266.62**

intermediate feature domain that is more compatible with optical imagery. Such feature-level alignment explains the substantial improvements in PSNR, SSIM, and MSE, and provides further evidence that the proposed translation model reduces cross-modal discrepancies not only at the pixel level but also in the learned feature space.

On the more challenging winter subset, where seasonal degradation further enlarges the modality gap, DiffOS achieves a PSNR gain of 5.42 dB, increases SSIM from negative correlation to 0.2823, and reduces MSE by approximately 70%. The performance advantage over pix2pix and CycleGAN becomes more pronounced under this adverse setting, indicating stronger robustness to extreme modality and seasonal variations. On OSDataset, direct SAR–optical similarity is very limited. DiffOS improves PSNR by 6.02 dB and SSIM by more than five times, while reducing MSE by 75.6%, demonstrating superior generalization under degraded optical conditions and large modality gaps. Overall, these results confirm that the proposed diffusion-based translation provides a more favorable representation for downstream matching.

C. Ablation Studies

1) Attentions: To analyze the contributions of hierarchical attention, ablation studies are conducted by selectively removing CCA, CKA, and CTA. Results on SEN1-2 and OSDataset are reported in Table IV. On the summer subset, removing either CCA or CKA causes moderate performance degradation, while removing both leads to a substantial drop in mIoU. This indicates that CCA and CKA provide complementary benefits. On the winter subset, performance degradation becomes more severe, highlighting the importance of robust feature interaction under strong appearance variations. On OSDataset, although absolute performance is lower, the full model consistently achieves the best balance between mean IoU and high-IoU ratio, demonstrating improved robustness under challenging conditions.

2) Losses: Loss ablation results are summarized in Table V. Removing either the local or global loss degrades performance across all datasets. The local loss plays a critical role in preserving fine-grained geometric alignment, especially under severe appearance distortions, while the global loss stabilizes overall matching behavior. Their combination yields the best and most consistent performance.

TABLE IV: The attention ablation study on the SEN1-2 and OSdataset.

(a) Summer subset

Att.	CCA	CKA	CTA	mIoU	IoU$_{>0.9}$(%)
CCA	✓			0.7958	58.56
CKA		✓		0.8283	73.97
	✓	✓		0.8503	78.77
CTA	✓	✓	✓	**0.8616**	**79.45**

(b) Winter subset

Att.	CCA	CKA	CTA	mIoU	IoU$_{>0.9}$(%)
CCA	✓			0.5781	23.00
CKA		✓		0.5997	25.67
	✓	✓		0.6285	27.33
CTA	✓	✓	✓	**0.6526**	**29.67**

(c) OSdataset

Att.	CCA	CKA	CTA	mIoU	IoU$_{>0.9}$(%)
CCA	✓			0.3251	0.00
CKA		✓		0.3299	2.16
	✓	✓		0.3996	**4.32**
CTA	✓	✓	✓	**0.4239**	2.88

TABLE V: The loss ablation study on SEN1-2 and OSdataset.

(a) Summer subset

Loss	Local	Global	mIoU	IoU $_{>0.9}$(%)
Local	✓		0.6994	49.32
Global		✓	0.8445	78.42
All	✓	✓	**0.8616**	**79.45**

(b) Winter subset

Loss	Local	Global	mIoU	IoU $_{>0.9}$(%)
Local	✓		0.5716	23.00
Global		✓	0.6092	**36.00**
All	✓	✓	**0.6526**	29.67

(c) OSdataset

Loss	Local	Global	mIoU	IoU $_{>0.9}$(%)
Local	✓		0.3645	2.08
Global		✓	0.4046	2.16
All	✓	✓	**0.4239**	**2.88**

3) Bidirectional Matching Performance: Bidirectional matching results are reported in Tables VI. On SEN1-2, DiffOS achieves high localization accuracy in both forward and reverse directions, with near-symmetric performance on the summer subset. On the winter subset, reverse matching slightly outperforms forward matching, suggesting that real optical patches provide more stable geometric cues under severe appearance degradation.

On OSDataset, although absolute accuracy is lower, DiffOS maintains consistent bidirectional performance, demonstrating robustness under challenging real-world conditions.

D. Comparison with State-of-the-Art Methods

DiffOS is compared with representative handcrafted and learning-based methods, including SIFT, TFeat [18], HardNet [9], and MatchosNet [19]. Results are shown in Tables VII and VIII.

DiffOS significantly outperforms all baseline methods on both SEN1-2 and OSDataset. While direct matching ap-

TABLE VI: Bidirectional matching performance of DiffOS on the SEN1-2 dataset and OSDataset.

Subset	Direction	Mean IoU	IoU > 0.9 (%)
Summer	Forward	**0.8616**	79.45
	Reverse	0.8528	**80.14**
Winter	Forward	0.6526	29.67
	Reverse	**0.6822**	**39.67**
OSDataset	Forward	0.4239	2.88
	Reverse	**0.4394**	2.16

TABLE VII: Comparison with state-of-the-art methods on Summer and Winter subsets.

Method	Summer	Winter
SIFT	0.3598	0.2755
TFeat	0.3311	0.2994
HardNet	0.3281	0.3363
MatchosNet	0.4921	0.3026
DiffOS	**0.8616**	**0.6526**

proaches suffer substantial degradation under cross-modal and seasonal variations, DiffOS maintains strong performance due to its decoupled translation-and-matching design. These results demonstrate the effectiveness of explicitly reducing modality gaps prior to correspondence estimation.

V. Conclusion

In this paper, we propose DiffOS, a unified framework for cross-modal SAR–optical image translation and matching. Unlike conventional pipelines that treat translation and matching as independent stages, DiffOS adopts a task-driven design in which cross-modal translation is explicitly optimized to facilitate downstream correspondence estimation. By integrating a conditional diffusion-based translation model with a dedicated matching network, the proposed framework effectively reduces the severe modality gap between speckle-dominated SAR imagery and optical images.

Extensive experiments on the SEN1-2 dataset and the more challenging OSDataset demonstrate that DiffOS consistently improves cross-modal matching accuracy under diverse and adverse conditions. Compared with classical handcrafted descriptors and recent learning-based matching methods, DiffOS achieves substantially superior localization performance. These results indicate that aligning cross-modal translation with the geometric requirements of matching provides a more effective and robust solution for SAR–optical correspondence estimation, and suggests a promising direction for future research on tightly coupled translation and matching in cross-modal vision tasks.

TABLE VIII: Comparison with state-of-the-art methods on OSDataset.

Method	Mean IoU	IoU > 0.9 (%)
TFeat	0.3242	0.72
HardNet	0.3407	1.44
SIFT	0.3303	1.62
MatchosNet	0.3701	2.88
DiffOS	**0.4239**	**2.88**

References

[1] X. Chen, Z. Dong, Z. Zhang, C. Tu, T. Yi, and Z. He, "Very high resolution synthetic aperture radar systems and imaging: A review," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 17, pp. 7104–7123, 2024.

[2] G. Zhao, J. Jiang, G. Dong, and H. Liu, "Sar ship detection via knowledge transfer: From optical image," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 18, pp. 17 526–17 538, 2025.

[3] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[4] M. Brown and D. G. Lowe, "Automatic panoramic image stitching using invariant features," *International Journal of Computer Vision*, vol. 74, no. 1, pp. 59–73, 2007.

[5] H. Bay, T. Tuytelaars, and L. Van Gool, "Surf: Speeded up robust features," in *European Conference on Computer Vision (ECCV)*, 2006, pp. 404–417.

[6] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "Orb: An efficient alternative to sift or surf," in *ICCV*, 2011, pp. 2564–2571.

[7] J. Liu, Q. Wang, D. Liao, J. Huang, T. Lai, and H. Huang, "Lpsd: Log-polar sampling descriptor for efficient and robust multimodal image matching," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 63, pp. 1–19, 2025.

[8] D. DeTone, T. Malisiewicz, and A. Rabinovich, "Superpoint: Self-supervised interest point detection and description," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2018, pp. 337–33 712.

[9] A. Mishchuk, D. Mishkin, F. Radenovic, and J. Matas, "Working hard to know your neighbor's margins: Local descriptor learning loss," in *NeurIPS*, 2017.

[10] J. Sun, Z. Shen, Y. Wang, H. Bao, and X. Zhou, "Loftr: Detector-free local feature matching with transformers," in *CVPR*, 2021.

[11] P. Lindenberger, P.-E. Sarlin, and M. Pollefeys, "Lightglue: Local feature matching at light speed," in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 17 581–17 592.

[12] Q. Wang, J. Zhang, K. Yang, K. Peng, and R. Stiefelhagen, "Matchformer: Interleaving attention in transformers for feature matching," in *Proc. Asian Conf. Comput. Vis. (ACCV)*, 2022, pp. 2746–2762.

[13] S. Li, G. Dong, and Y. Fan, "Focusnet: A new data-driven sar image autofocus method," in *Proceedings of the International Joint Conference on Neural Networks (IJCNN)*, 2025, pp. 1–6.

[14] S. Li and G. Dong, "Ifenet: Sar imaging on the missing echo via image formation and enhance network," in *Proceedings of the International Joint Conference on Neural Networks (IJCNN)*, 2025, pp. 1–8.

[15] M. Schmitt, L. H. Hughes, and X. X. Zhu, "The SEN1-2 dataset for deep learning in SAR-optical data fusion," *CoRR*, vol. abs/1807.01569, 2018.

[16] Y. Xiang, X. Wang, F. Wang, H. You, X. Qiu, and K. Fu, "A global-to-local algorithm for high-resolution optical and sar image registration," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–20, 2023.

[17] X. Yuming, C. Jinyang, and H. Zhonghua, "Osdataset2.0: Sar-optical image matching dataset and evaluation benchmark," *Journal of Radars*, 2025.

[18] V. Balntas, E. Riba, D. Ponsa, and K. Mikolajczyk, "Learning local feature descriptors with triplets and shallow convolutional neural networks," in *British Machine Vision Conference (BMVC)*, 2016.

[19] Y. Liao, X. Huang, H. Liu, and L. Zhang, "Matchosnet: A deep learning network for sar and optical image patch matching," *IEEE Geoscience and Remote Sensing Letters*, vol. 16, no. 10, pp. 1551–1555, 2019.