

Interpretable Matching of Optical-SAR Image via Dynamically Conditioned Diffusion Models

Shuiping Gou
shpgou@mail.xidian.edu.cn
Xidian University
Xi'an, China

Xinlin Wang*
wangxinlin@xidian.edu.cn
Xidian University
Xi'an, China

Xin Wang
23171214441@stu.xidian.edu.cn
Xidian University
Xi'an, China

Yunzhi Chen
2016010026@hzvtc.edu.cn
Hangzhou Vocational and Technical College
Hangzhou, China

1 PRELIMINARIES

Most diffusion models are based on the framework of DDPMs [1], which consist of a forward diffusion process and a reverse generation process, both of which are modelled as Markov chains. During forward diffusion, Gaussian noise is slowly added to the original data x_0 according to a fixed variance schedule $\{\beta_1, \beta_2 \dots \beta_T\}$ in T steps. Finally, a sequence of noisy samples $\{x_0, \dots, x_t, \dots, x_T\}$ is produced, where $x_T \sim N(0, \mathbf{I})$. Each diffusion step is formulated as:

$$q(x_t | x_{t-1}) := N(x_t; \sqrt{1 - \beta_t}x_{t-1}, \beta_t \mathbf{I}) \quad (1)$$

By setting the $\alpha_t = 1 - \beta_t$ and the $\bar{\alpha}_t = \prod_{i=1}^t \alpha_i$, x_T can be obtained at any moment t by the following equation:

$$q(x_t | x_0) := N(x_t; \sqrt{\bar{\alpha}_t}x_0, (1 - \bar{\alpha}_t) \mathbf{I}) \quad (2)$$

During the reverse generation process $p(x_{t-1} | x_t)$, the noise is gradually removed and the original data is reconstructed. To solve the problem, the DDPM learns the parametric Gaussian transform $p_\theta(x_{t-1} | x_t)$. Essentially, it predicts the mean of the Gaussian distribution $\mu_\theta(x_t, t)$. The reverse process is expressed as follows:

$$p_\theta(x_{t-1} | x_t) = N(x_{t-1}; \mu_\theta(x_t, t), \sigma_t^2 \mathbf{I}) \quad (3)$$

During training, the denoising neural network $f_\theta(x_t, t)$ is trained to minimize the training objective via L2 loss, that is, predicting x_0 from the x_t [1]:

$$L = \|f_\theta(x_t, t) - x_0\|^2 \quad (4)$$

In the inference phase, data samples x_0 are reconstructed from the noise x_T in an iterative manner using the model f_θ and updating rules [1].

2 ADDITIONAL EXPERIMENTAL RESULTS

2.1 The universality of interpretable latent space

The interpretable latent space proposed is derived from an auto-encoder pre-trained on the OSdataset. However, SAR images from different remote sensors exhibit distinct characteristics. To demonstrate the matching generality of this interpretable latent space, as depicted in Figure 1, inference was conducted on images from both the OSdataset and the SEN1-2 datasets. The translated images (Template \hat{S}), translated features (Template \hat{O}'), and similarity

maps were visualized. For samples from the OSdataset, the translated template images and features closely resemble the actual template images (Template S) and features (Template S'), yielding correct matching results. Conversely, on the diverse SEN1-2 datasets, the translated SAR templates exhibit reduced overall brightness compared to the ground truth SAR templates, owing to the higher overall brightness in the OSdataset. Nevertheless, this disparity does not affect the matching. Moreover, the SAR images in the SEN1-2 datasets display more noise than those in the OSdataset. Despite this, the translated images and features exhibit less noise while preserving the structural connectivity features. Furthermore, the resulting similarity maps continue to exhibit smooth single peaks, precisely located within the correct matching regions. These observations collectively underscore the generality of the auto-encoder trained on the OSdataset, representing a one-time investment.

2.2 Visualisation of the impact of the ACB

The proposed Attention Calibration Block (ACB) operates on 2D attention maps following cross-attention. To assess the smoothness and calibration efficacy of our proposed ACB on attention maps, as depicted in Figure 2, we visualize the ground truth, pre-correction, and post-correction attention maps of four samples during the inference process on two datasets. Specifically, we selected an urban built-up area and a rural cropland area for each dataset, with the rural farmland area posing greater matching difficulty due to its numerous highly similar areas. Samples (a) and (c) are from the OSdataset, while samples (b) and (d) are from SEN1-2. The sampling steps were set to 5, and at the final sampling iteration, we visualize the attention map before and after ACB with a feature stride of 8.

Observing the attention maps, it is evident that they exhibit diagonal sparsity. For samples (a) and (b) from the urban built-up area, cross-attention accurately perceives rich texture details; however, the attention maps display numerous points with inconsistent values and a few with incorrect locations. Following ACB, the corrected attention maps no longer contain tokens with incorrect locations and are smoothed. It is noteworthy that while ACB effectively calibrates the attention map, a slight inconsistency with the ground truth is observed. Nonetheless, our translation results still yield correct matches. This is because although the attention map contains information about the match location, refining the attention region step by step is necessary, and direct matching using

*Corresponding author.

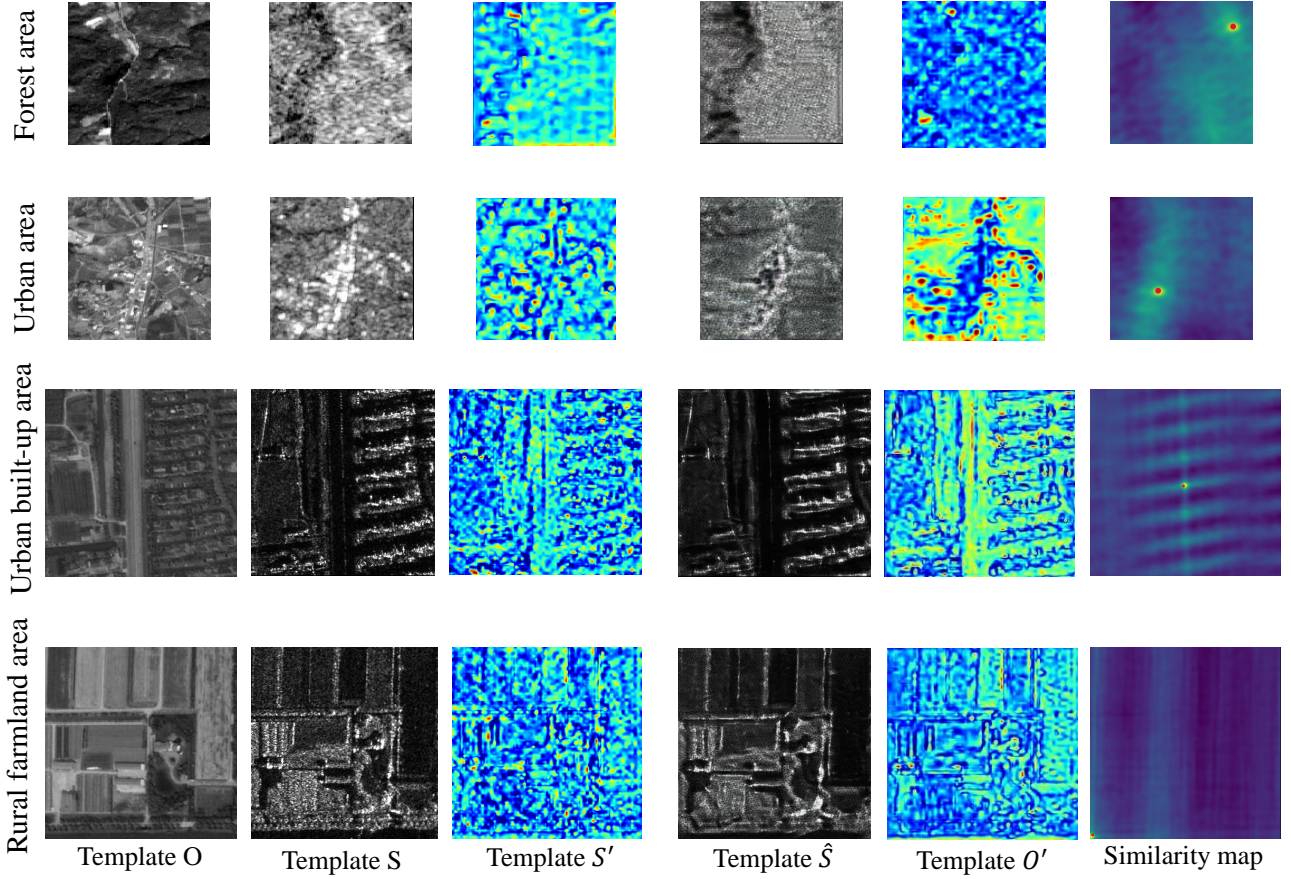


Figure 1: Visualisation of translated SAR template images (Template \hat{S}), the translated features (Template O'), and the similarity maps from different areas of the OSdataset and SEN1-2 datasets. The auto-encoder trained on the OSdataset possesses a certain degree of universality, which is a one-time investment.

the attention map is suboptimal compared to matching using the translated image.

Rural farmland areas comprise numerous similar rectangular regions, often resulting in erroneous perceptions by cross-attention. In samples (c) and (d), it is evident that the values and locations of most tokens in the uncalibrated attention map are confusing and inconsistent. However, surprisingly, the ACB module can still consolidate all the information from the attentional graph to produce a more consistent attentional graph. Calibration in rural farmland areas is slightly inferior compared to urban built-up areas due to the challenges associated with matching in rural farmland areas. These findings demonstrate that the proposed attention calibration block possesses effective local calibration capabilities and some global calibration capabilities, leading to more accurate attention perception and matching effects.

2.3 Analysis of Interpretability

Existing methods extract common features of multimodal images for matching, which is abstract, and hard to understand its similarity. However, our approach is built on cross-modal translation, which translates one modality into the other modality in the latent space, and completes matching in the translated feature space. We adopt structural similarity index measure (SSIM) to constrain the translation process. Figure 3 (a) visually shows the matching feature similarities of SAR-optical images. It shows that our model can capture similar structures (red circle) via modality translation, whereas common cross-modal features extracted by MARU-Net are inconsistent. Figure 3 (b) quantitatively shows that the more similar the features, the better the matching.

2.4 Analysis of Robustness

Most methods extract unknown common features of multimodal images, which is vulnerable to the interference. Our translation-based method focuses on consistency structural features, which is

insensitive to noises. Figure 4 visually compares the correct matching rate (CMR) [2, 3] of SOTA methods on noisy optical template images in the OSdataset without further training. It shows that the CMR of all methods decreases with the addition of more noises. However, the CMR of our model only has a very slight drop when the noise variance is less than 5%, and consistently remains higher than that of others on different noisy data.

2.5 Analysis of Generalization

In the submission, the experimental datasets, OSdataset and SEN1-2, consist of optical-SAR image pairs. To evaluate the generalization, we conduct experiments on an extra NIR (Near Infrared)-SAR dataset containing 100 NIR-SAR image pairs with a size of 5556×3704 pixels and 5m spatial resolution. These large-scene images are crop into 5521 image pairs with the size of 512×512 . The NIR images are treated as templates, with the training, validation and testing datasets split in a 7:2:1 ratio. Firstly, we trained SOTA models on the NIR-SAR dataset (w/ Training) to evaluate each approach. In addition, we tested the NIR-SAR testing dataset directly using SOTA models trained on the OSdataset (w/o Training) to further assess the generalization. Table 1 quantitatively shows the generalization performance from different views.

Table 1: The comparison of SOTAs on the NIR-SAR dataset.

| Methods | w/ Training | | | w/o Training | | |
|------------|---------------|---------------|---------------|---------------|---------------|----------------|
| | CMR (T=5) | RMSE (T=5) | RMSE (all) | CMR (T=5) | RMSE (T=5) | RMSE (all) |
| OSMnet | 0.7852 | 2.7871 | 13.2734 | 0.5917 | 3.1245 | 16.1871 |
| MARU-Net | 0.8129 | 2.4707 | 10.1147 | 0.6226 | 2.6082 | 15.9857 |
| RSOMNet[3] | 0.8160 | 2.1015 | 8.9358 | 0.6832 | 2.3044 | 14.0608 |
| Ours | 0.8341 | 2.2312 | 8.7958 | 0.7188 | 2.3874 | 11.2455 |

Table 2: The performance of SOTA methods on two datasets.

| Methods | OSdataset | | | | SEN1-2 | | | |
|------------|-----------|-----------|-----------|----------|----------|-----------|-----------|---------|
| | CMR(T=5) | RMSE(T=5) | RMSE(all) | Time(s) | CMR(T=5) | RMSE(T=5) | RMSE(all) | Time(s) |
| RSOMNet | 0.8296 | 2.0834 | 7.9752 | 0.1847 | 0.9108 | 1.4793 | 5.0752 | 0.1021 |
| VSMATCH | 3.06 | 2.2e5 | 0.86 | 451.2569 | 3.06 | 5.5e4 | 0.52 | 87.155 |
| OSMnet | 5.53 | 42.92 | 1.29 | 0.0913 | 5.53 | 11.2 | 0.67 | 0.0568 |
| MARU-Net | 18.82 | 39.6 | 1.14 | 0.0895 | 18.82 | 10.3 | 0.64 | 0.0346 |
| RSOMNet | 13.57 | 140.9 | 3.24 | 0.1847 | 13.57 | 21.6 | 1.20 | 0.1152 |
| Ours (w/o) | 20.7+5.6 | 61.7 | 2.09 | 0.1572 | 20.7+5.6 | 13.4 | 1.01 | 0.1021 |
| Ours (w/) | 21.1+5.6 | 51.1↓ | 1.45↓ | 0.1093 | 21.1+5.6 | 12.6↓ | 0.74↓ | 0.0621 |

Pars: Parameters Com: Computational complexity Mem: Memory usage

2.6 Model Performance

Figure 5 visually compares SOTA models on SEN1-2 dataset, which displays the superiority of our method. Table 2 provides a quantitative comparison of an advanced matching method published in 2024, RSOMNet[3]. It is observed that our method performs better than RSOMNet on all metrics on the SEN1-2 dataset. For the OSdataset, our approach (RMSE(T=5)=2.29) is inferior to RSOMNet (RMSE(T=5)=2.0834) on the RMSE(T=5). The OSdataset is a high-resolution dataset, which contains richer details. However, our translation-based method focuses more on the structural features. Therefore, our method is not state-of-the-art on the pixel-level metric RMSE, but shows promising match accuracy. ‘

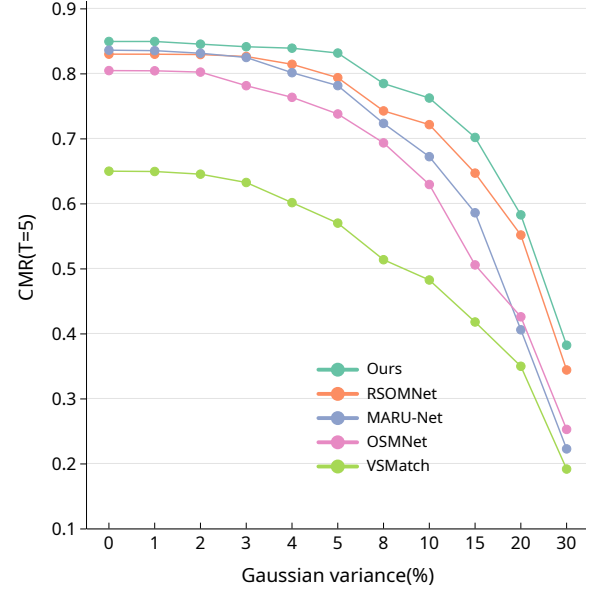


Figure 4: CMR of SOTAs on noising OSdataset.

2.7 Model Parameters

Table 2 further details parameters, computation, and memory usage of several SOTA methods during inference. For the large-scale OSdataset, the sparse attention mechanism (Ours(w/)), compared to the method without sparse attention (Ours(w/o)), significantly reduces memory and computation resource usage while only increasing parameters by 1.5%. Our model consists of both trainable and frozen parameters (the encoder), where trainable parameters are comparable to MARU-Net’s. RSOMNet requires very large computation resources.

REFERENCES

- [1] Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems* 33 (2020), 6840–6851.
- [2] Yifan Wang, Xingyi He, Sida Peng, Dongli Tan, and Xiaowei Zhou. 2024. Efficient LoFTR: Semi-dense local feature matching with sparse-like speed. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 21666–21675.
- [3] Hong Zhang, Yuxin Yue, Haojie Li, Pan Liu, Yusheng Jia, Wei He, and Zhihui Wang. 2024. Shared contents alignment across multiple granularities for robust SAR-optical image matching. *Information Fusion* 106 (2024), 102298.

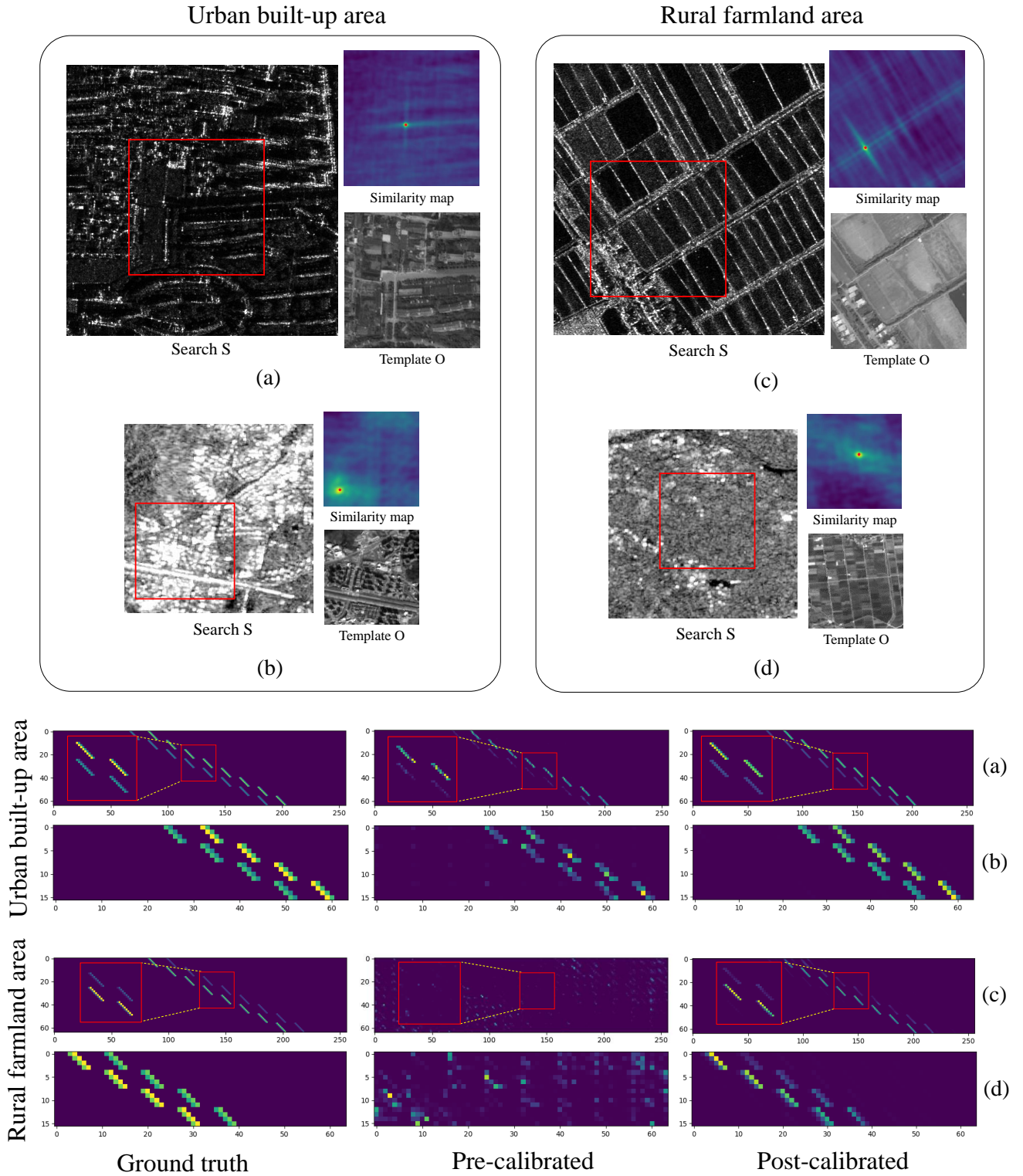


Figure 2: Visualisation of ground truth, pre-calibrated, and post-calibrated attention maps on 4 samples. The two samples (a) and (b) are from urban built-up areas, and the samples (c) and (d) are from Rural farmland areas. Samples (a) and (c) are from the OSdataset and samples (b) and (d) are from the SEN1-2 datasets. These results show that the proposed attention calibration block has an effective local calibration capability and some global calibration capability.

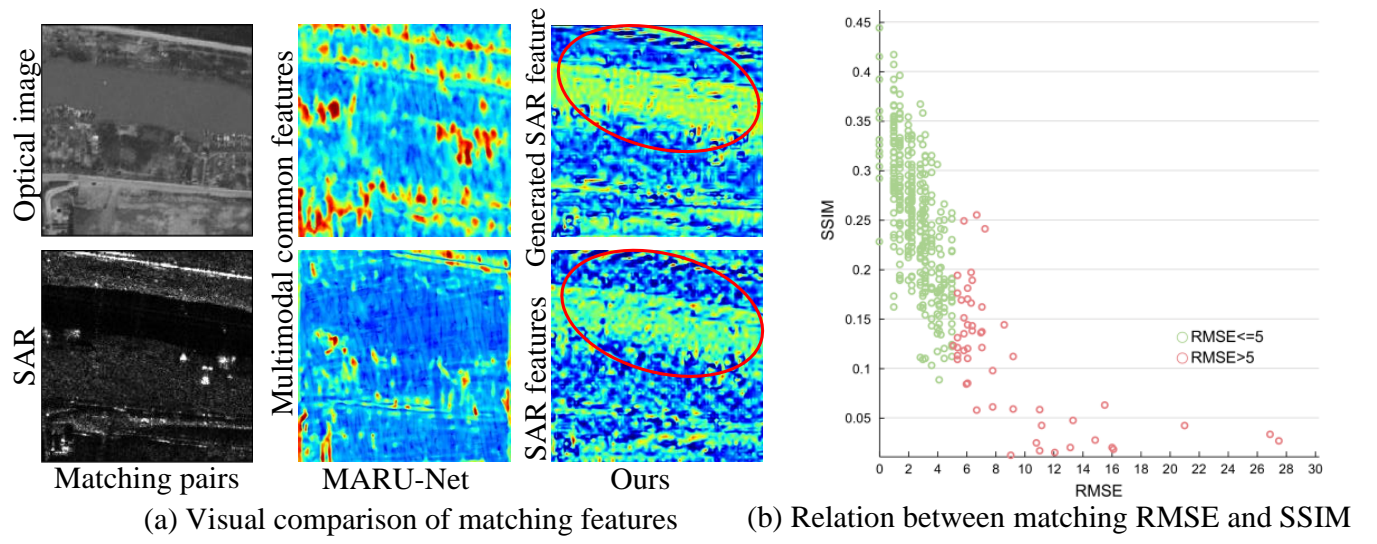


Figure 3: The interpretability explanation.

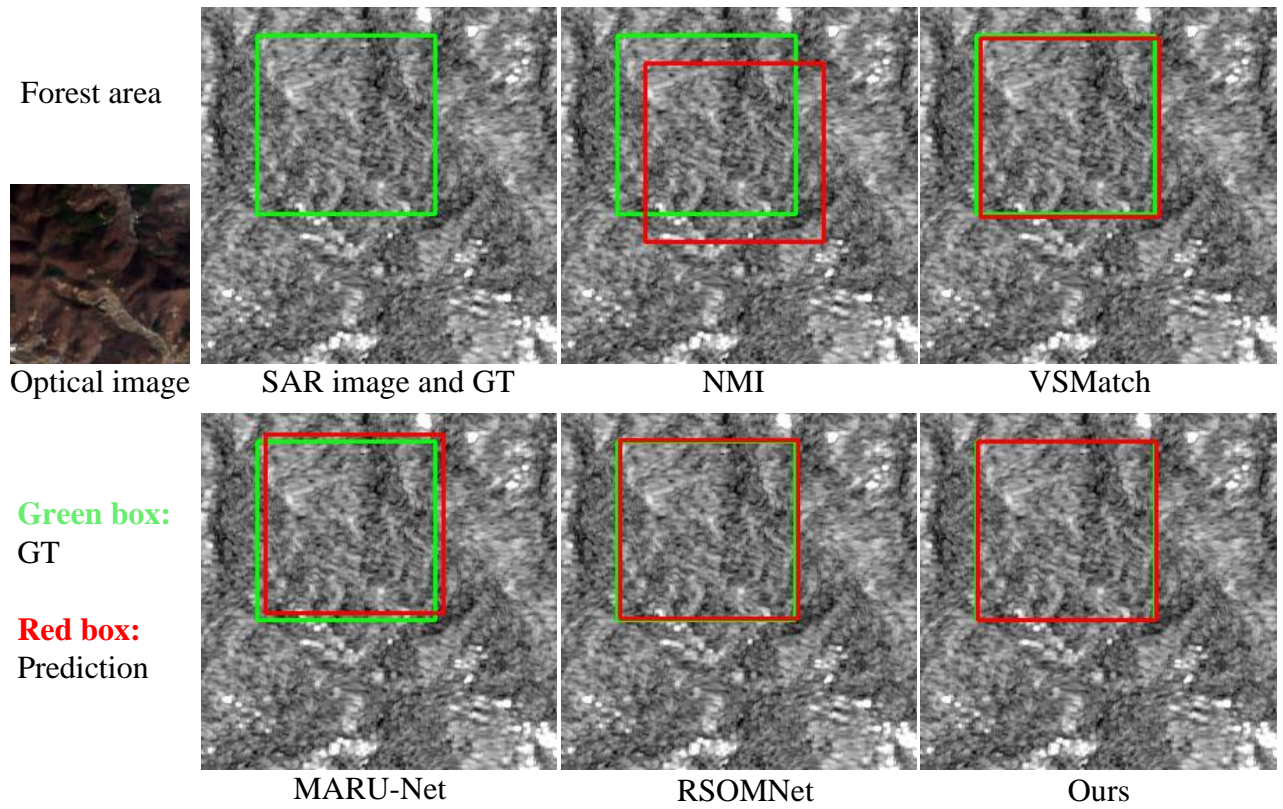


Figure 5: The visualization of SOTAs on SEN1-2 dataset.