HOMO-Feature: Cross-Arbitrary-Modal Image Matching with Homomorphism of Organized Major Orientation

Supplementary Material

1. Details of the proposed GCZ dataset

The General Cross-modal Zone (GCZ) dataset is a multi-modal, one-scene, real-world dataset, designed specifically for the purpose of cross-arbitrary-modal image matching. It is a multi-source remote sensing imaging data, encompassing a wide range of modalities, including RGB, PAN, NIR, SWIR, MSI, HSI, SAR, LiDAR, Map, and others. The overall data is visualized in Fig. 1. The scene includes a variety of ground features, such as land, water bodies, and coastlines. The land features include plains, mountains, hills, forests, bare rocks, etc., while the anthropogenic objects include farmland, buildings, roads, bridges, and man-made canals. The aim is to incorporate a variety of elements with their potentially unique textures and structural characteristics to thoroughly demonstrate the algorithm's efficiency.

Through testing, we find that the optical modalities are similar in morphology, resulting in excellent matching performance. However, this similarity makes it difficult to fully test the cross-modal capabilities of HOMO. Therefore, in this experiment, we simplify GCZ to six major categories: VIS, IR, SAR, Depth, Map, and Other, which are sufficient to test the integrated capabilities of the algorithms. The total size of GCZ is 5780×22000. Patches are standardized using a 4096×4096 cropping window, with a slightly more than 50% overlap, and then resized to 512×512. And in the experiments, using the randomly distorted sample creation method shown in Fig. 2 and filtering out invalid samples, a total of 4752 samples are obtained. This makes the GCZ be the largest publicly available, real-world, generally cross-modal image matching dataset.

This dataset has been manually matched, calibrated, and reviewed carefully, resulting in a standardized registered groundtruth that can be used for distorted sample generation and training data creation. The dataset will be released in the form of processed three-channel RGB-visualized PNG images, mainly for researching purpose on image matching.

2. Details of fine-tuning of compared networks

The training samples of GCZ are created using the method shown in Fig. 2, which is also widely adopted in many related works. In the process, patches are randomly transformed until the training (tuning) ends. Fine-tuning is performed on D2-Net and LoFTR, as two representative network frameworks of detector-based CNN and detector-free CNN+Transformer. The labels, loss, and other details are kept consistent with the official implementation. 30% of randomly selected cropped GCZ samples are used as training set, utilizing the distortion method in Fig. 2 until the training ends.

However, with training of randomly distorted samples, these models could not handle such wide ranges of cross-modal samples with arbitrary rotations. The training ultimately resulted in these networks being effective either only within a small range of rotation angles ($\pm15°$) for each modality, or being invariant to any rotation but only under specific modalities. This revealed the inherent limitations of these models, indicating that larger models must be utilized. The number of fine-tuning epoch are 60 and 10 for D2-net and LoFTR in the experiments, respectively, which produces the best performances.

3. Parameters of HOMO

The fixed value of all settable parameters of HOMO in the experiments are listed in Tab. 1. All parameters are selected based on extensive parameter sweeps to achieve the best overall performance. It is worth noting that the downsampling ratio is set to 1.2, with an octave number of 4, ensuring the ability to accommodate scale differences up to a ratio of $1.2^4 > 2$. Larger scale differences in practical data can be easily handled by pre-sampling the images to an appropriate range. The maximum points number in DoM keypoint detection is set to 3600 to ensure fairness with most detector-free network methods under 512×512 input samples, based on the calculation: $(\text{size}/\text{coarse-boundary})^2 \rightarrow (512/8-4)^2 = 3600$. All other compared methods use their original default parameters.

Table 1. Pre-fixed parameters of the HOMO.

Module	Parameter	Value
MOM	LG scale number	4
	LG orientation number	6
	ASW scale number	4
DoM	LNMS window size	3
	Maximum points number	3600
GPolar	Structure size R_2	72
	Cell division N_A	12
	Angle division N_O	12
MsS	Octave number N_octave	4
	Layer number N_layer	3
	Downsampling ratio	1.2
	Gaussian blurring σ	1.6

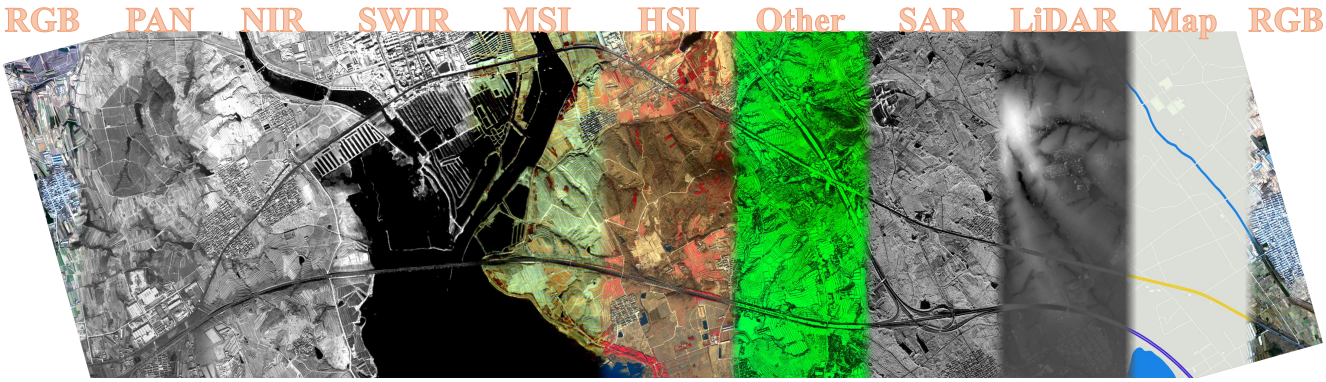

Figure 1. The visualization overview of our proposed GCZ dataset.

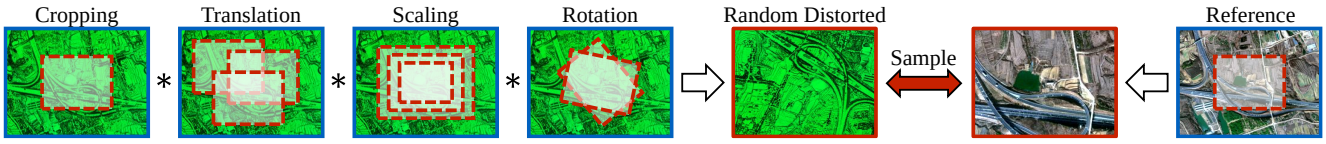

Figure 2. Creation method of randomly distorted GCZ samples.

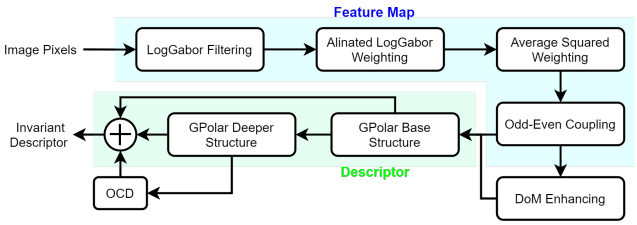

Figure 3. Feature flow of the input image in the whole HOMO framework.

4. Ablation study

The proposed HOMO framework consists of a series of feature extraction and transformation processes. We summarize the core processing steps in Fig. 3. The original image undergoes a sequence of modules, resulting in the feature flow illustrated in the figure. This process is regarded as a complete transformation from the image pixel space to an invariant feature space. Based on this structure, we conducted detailed ablation tests by removing or replacing key modules with other typical components. The complete ablation testing results are presented in Tab. 2.

Specifically, "MOM $\rightarrow$ Gradient [0,2π)" refers to degrading the MOM values, which represent gradient-like orientations, into classical gradients computed using the Sobel operator—essentially the basic features used in SIFT and related algorithms. "MOM $\rightarrow$ Gradient [0,π)" indicates enforcing the gradient orientations to lie within [0,π), by mapping values outside this range to their opposite directions, as used in PIIFD and related algorithms. "Gabor $\rightarrow$ Gradient" means replacing the LogGabor filtering step in the MOM with gradient computation, while keeping subsequent steps such as ASW unchanged.

When replacing the GPolar with other descriptors, only the spatial structure of the descriptor is changed, while the feature statistics approach remains the same, still based on MOM values. The reference direction calculation and the OCD module are also preserved, in order to independently evaluate the effectiveness of the proposed descriptor structure.

The results are analyzed in the main paper.

Table 2. Ablation study on portable modules of HOMO with average NCM under each rotation angle. "w/o" indicates "without".

Average NCM under rotation		30°	45°	90°	120°	180°
	HOMO	**1367**	**1331**	**1854**	**1367**	**1856**
MOM	$\rightarrow$ Gradient [0,2π)	599	313	462	134	896
	$\rightarrow$ Gradient [0,π)	1072	1037	1499	1032	1525
	Gabor $\rightarrow$ Gradient	934	906	1350	862	1402
	w/o ASW	1030	1005	1406	1032	1408
	w/o even-ASLG	1358	1320	1717	1365	1723
	w/o DoM	1279	1142	1851	1366	1851
GPolar	$\rightarrow$ HOG (square)	971	896	1009	891	991
	$\rightarrow$ GLOH (polar)	674	732	990	679	990
	$\rightarrow$ LogPolar (polar)	1222	1248	1700	1224	1699
	w/o deeper structure	1365	1329	1852	1367	1851
	w/o OCD	1301	1236	1518	958	64
	w/o OCD and deeper	1156	1035	1103	699	21

Figure 4. Failure cases.

5. More visual results

A more comprehensive visualization of the image matching results on the proposed GCZ dataset is shown in Fig. 5. In addition, we present some representative matching examples on the original MRSI dataset in Fig. 6. Since the samples in MRSI exhibit little to no rotational variation, most network-based algorithms perform well. To ensure a fair comparison, we adopted the original non-rotation version of RIFT (denoted as RIFT^{+}), and similarly removed the rotation operation in HOMO (denoted as HOMO^{+}).

It is worth noting that the MINIMA-tuned version of LoFTR (MINIMA$_{LoFTR}$) performs worse on the MRSI dataset compared to the version tuned on our GCZ dataset, even showing signs of negative optimization. This indirectly demonstrates the generalization capability and high value of the proposed GCZ dataset. Notably, HOMO, even when relying purely on handcrafted feature transformations, can still rival many data-driven network-based models, making it a strong candidate in scenarios with limited or unavailable training data.

6. Limitations and Failure cases

Failure cases of HOMO are illustrated in Fig. 4. These cases typically arise in samples with extremely sparse textures and contents, or in those with highly repetitive or blurry structures. This issue is particularly prominent in modalities such as map or day-night images. Such scenarios pose significant challenges for traditional methods—including HOMO—that rely on fixed local and low-level features. Addressing these challenges likely requires more global and deeper semantic features with broader receptive fields.

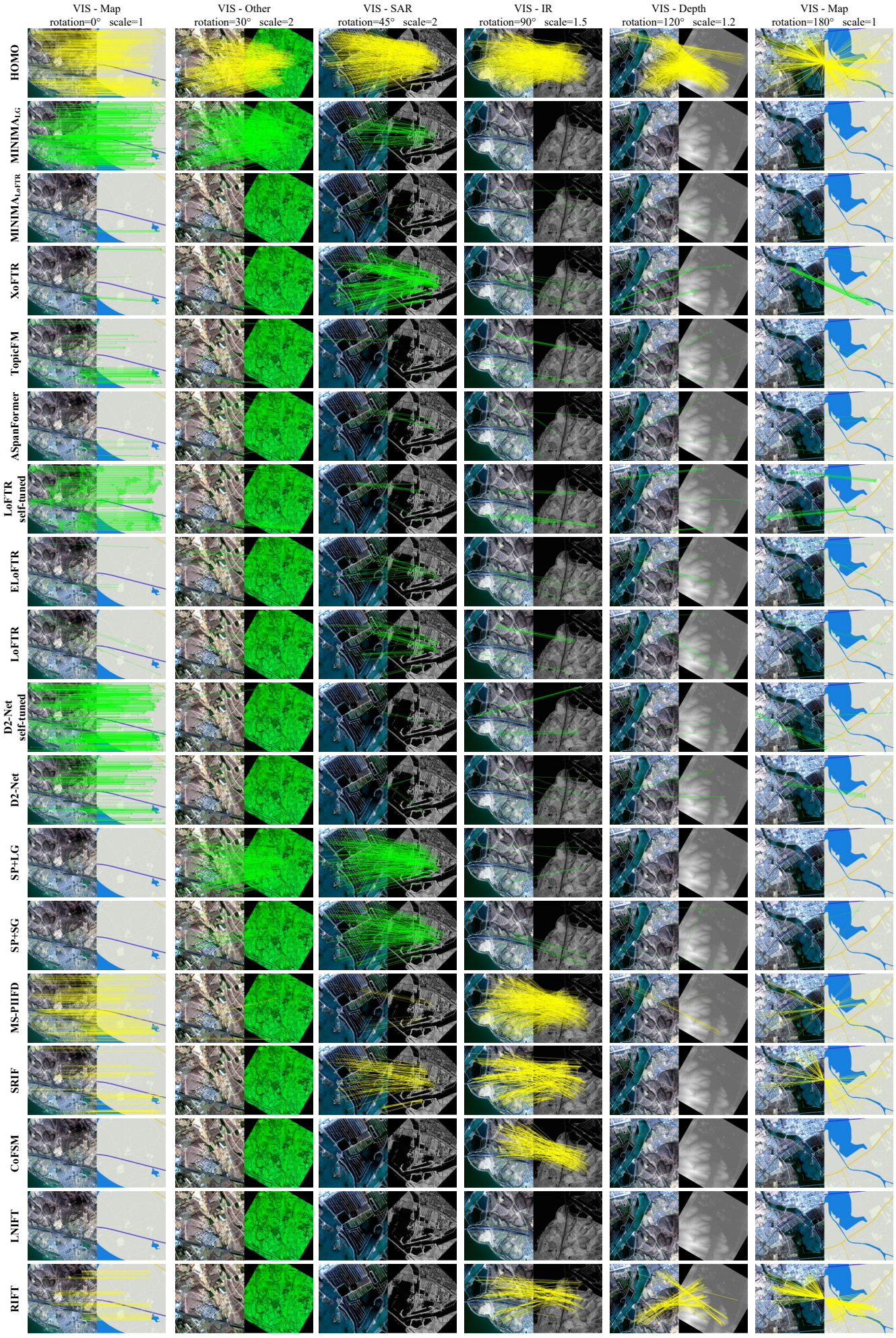

Figure 5. Examples of visual feature matching results under different simulated rotations and scale ratios on the proposed GCZ dataset. Matches after RANSAC are drawn. (**yellow**: traditional handcrafted methods, **green**: deep-learning network methods).

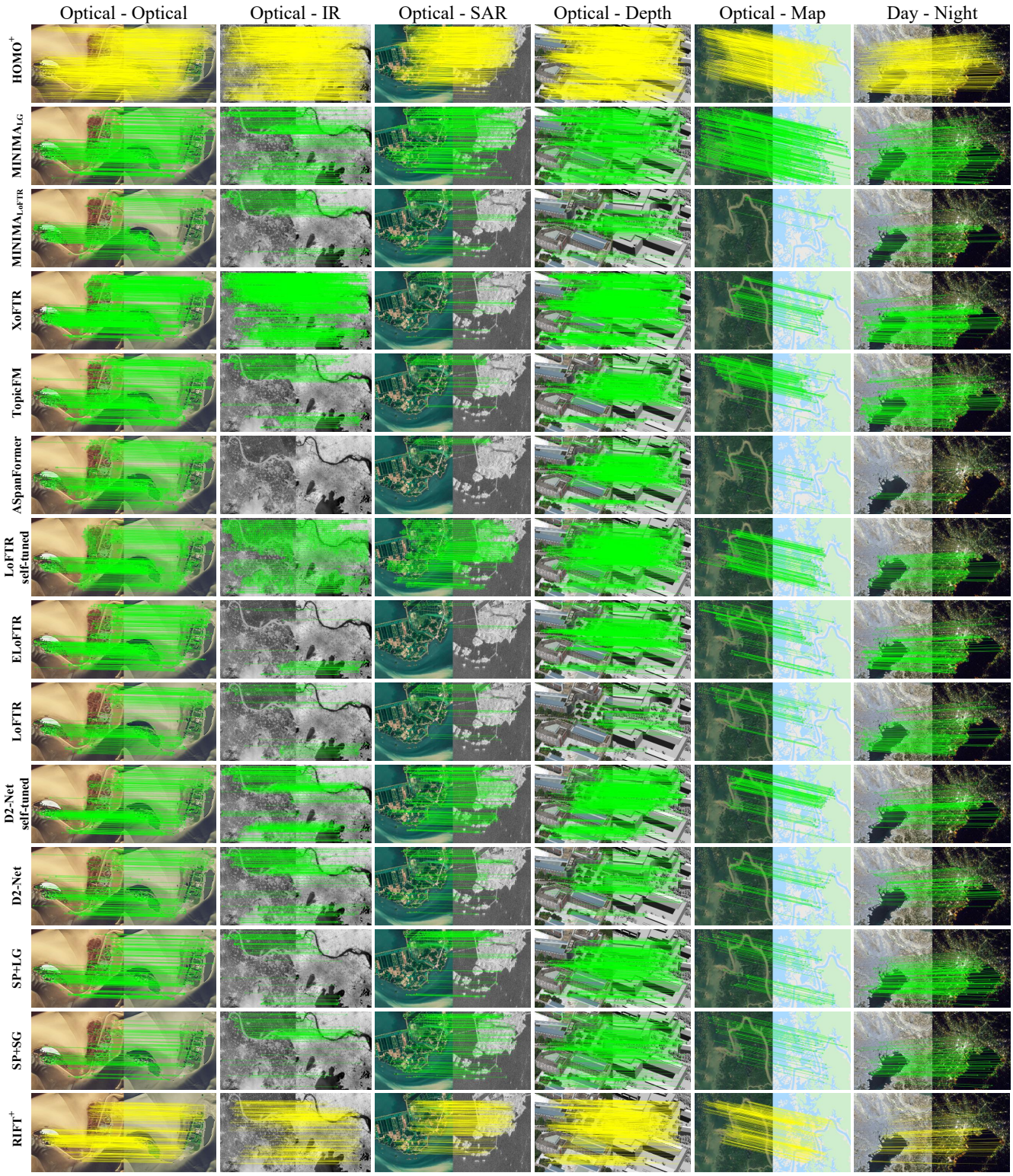

Figure 6. Examples of visual feature matching results on the MRSI dataset. Matches after RANSAC are drawn. (**yellow**: traditional handcrafted methods, **green**: deep-learning network methods).