SGPFeat: Semantic and Geometric Priors for Multi-modal Image Matching

Yuxin Deng[1*], Botian Wang[1*], Kaining Zhang[2], Hao Zhang[1†], Jiayi Ma[1†]

[1]Electronic Information School, Wuhan University, Wuhan 430072, China
[2]School of Artificial Intelligence and Robotics, Hunan University, Changsha 410082, China
{acuo.dyx, zkn707196, zhpersonalbox, jyma2010}@gmail.com, wangbotian@whu.edu.cn

Abstract

Multi-modal image matching is a fundamental task in multi-view and multi-modal image processing. Its key challenge lies in extracting features that remain consistent despite drastic appearance variations across modalities. However, the learning of the feature is hindered by the scarcity and the inaccurate alignment of existing multi-modal datasets. To address this, we propose a knowledge distillation framework termed SGPFeat that transfers rich prior knowledge from large-scale unimodal tasks to enhance multi-modal representation learning. Specifically, semantic priors from a vision foundation model guide the feature extractor to identify shared semantic structures across modalities, enabling better generalization under large appearance gaps. In parallel, geometric priors derived from accurately aligned visible-light datasets improve detection precision on noisy aligned multi-modal pairs. Furthermore, we introduce a Heterogeneous Feature Aggregation (HFA) module to facilitate effective distillation and feature representation. Extensive experiments demonstrate that semantic and geometric priors bring significant improvement for our SGPFeat across diverse multi-modal image matching benchmarks.

Code — https://github.com/wbt6661/SGPFeat

Introduction

Multi-modal image matching is a fundamental task in many cross-modal image processing applications, such as satellite-based disaster assessment (Li et al. 2022a), and computer-aided diagnosis in medical imaging (Guo et al. 2019). At the heart of this task lies feature extraction, which constitutes the preliminary and critical step in the matching process (Zhang et al. 2025b). Nevertheless, unlike in the visible domain, multi-modal feature extraction remains highly challenging due to substantial nonlinear appearance variations across sensing modalities (Jiang et al. 2021).

To handle these variations, hand-crafted methods (Li, Hu, and Ai 2019; Li et al. 2022b) detect common structures like edges and corners across modalities, using them as the basis for sparse feature detection and description. For example, RIFT (Li, Hu, and Ai 2019) detects keypoints on phase congruency maps (Kovesi 2003), which highlight stable edges and corners. And it further incorporates frequency information such as log-Gabor filters to enhance descriptor robustness against nonlinear distortions. However, such hand-crafted modeling still faces fundamental limitations in robustness and generalization. With the advent of deep learning, data-driven approaches have shown great potential by enabling neural networks to implicitly learn modality-invariant representations. Nevertheless, deep-learning-based methods face two key challenges (Jiang et al. 2021):

(i) The scarcity of aligned multi-modal image pairs. Unlike aligned visible-light image pairs, which can be easily obtained using off-the-shelf 3D reconstruction pipelines (Schonberger and Frahm 2016; Li and Snavely 2018), multi-modal pairs often require manual annotation. Although recent methods such as MINIMA (Ren et al. 2025) and MatchingAnything (He et al. 2025) attempt to synthesize training pairs by translating visible images into other modalities, generative models often introduce artifacts and distribution shifts, which degrade training accuracy. Consequently, the limited availability of real multi-modal data constrains the robustness of learned feature descriptors.

(ii) The inaccurate alignment in existing multi-modal datasets (Deng and Ma 2022; Jiang et al. 2021). These datasets are often annotated under simplified geometric assumptions such as affine or homography transforms, even though real cross-modal relationships frequently deviate from such transform. Such geometric inaccuracies lead to unreliable and imprecise feature detection, ultimately de-

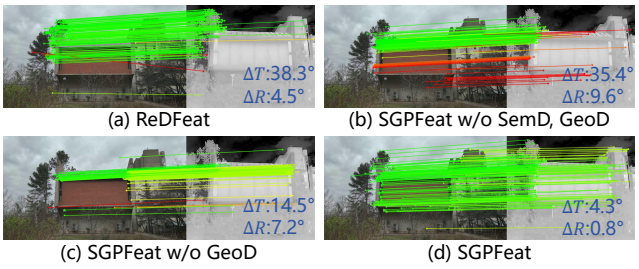

Figure 1: Visible-infrared image matching samples. SemD/-GeoD: semantic/geometric prior; $\Delta T/\Delta R$: the estimated relative translation/rotation error.

*These authors contributed equally.
†Corresponding author.

grading the overall matching performance. As illustrated in Fig. 1(a), the state-of-the-art method ReDFeat (Deng and Ma 2022), when trained on such misaligned data, tends to detect unstable and inaccurate keypoints in regions, such as the textureless sky, which is harmful to accurate image matching and relative camera pose estimation.

In this paper, we introduce two knowledge distillation techniques (Gou et al. 2021) to address the aforementioned challenges: *(i)* To enhance feature learning robustness under limited data, we distill rich semantic knowledge from a vision foundation model rather than generating large-scale synthetic datasets. Trained on diverse and extensive data, the vision foundation model possesses strong semantic understanding across varied scenes. By transferring this semantic prior via feature-level distillation, our model can capture shared semantics across modalities despite large appearance differences and sparse correspondences, thereby improving the robustness of feature descriptors for image matching, as illustrated in Fig. 1(c). *(ii)* To enhance the feature learning accuracy under inaccurate data, we distill accurate geocentric knowledge from a counterpart trained on large-scale and accurately aligned visible-light dataset. This geometric prior is transferred through label distillation, enabling accurate and reliable keypoint detection. In parallel, feature distillation transfers its geometric description knowledge, producing geometry-invariant descriptors. Together, these enhance the stability and accuracy of image matching.

Moreover, we develop a network architecture incorporating a Heterogeneous Feature Aggregation (HFA) module, which adaptively integrates multi-scale representations from Convolutional Neural Networks (CNNs) and Transformer blocks. This design facilitates modality-invariant feature learning and enables effective distillation from heterogeneous teacher models. Building upon this architecture and the fundamental constraints of ReDFeat, our proposed model, SGPFeat, seamlessly integrates both semantic and geometric priors, achieving outstanding performance across multiple cross-modal matching benchmarks.

Overall, we make the following contributions:

- We introduce a **semantic prior distillation** strategy that injects semantic knowledge from a vision foundation model to enhance descriptor robustness on limited multi-modal training data.
- We propose a **geometric prior distillation** method that transfers geometric consistency from a visible-image matching network to improve keypoint detection accuracy in noisy multi-modal datasets.
- We design a dedicated neural network architecture tailored for multi-modal feature learning and compatible with heterogeneous distillation.
- Extensive experiments on several cross-modal matching tasks demonstrate the robustness and accuracy of our proposed SGPFeat.

Related Work

Visible Image Matching

Image matching methods are typically categorized into three types (Zhang et al. 2025b): sparse (Mishchuk et al. 2017; Sarlin et al. 2020; Li, Zhang, and Ma 2024), semi-dense (Sun et al. 2021; Wang et al. 2024b), and dense matching (Edstedt et al. 2024b). Sparse methods rely on a limited set of keypoints, while semi-dense and dense methods operate on dense feature maps. Owing to their multi-view consistency, lower computational cost, and effectiveness in downstream tasks, sparse matching remains widely used across many image matching applications (Schonberger and Frahm 2016; Sarlin et al. 2019).

The sparse matching pipeline consists of feature extraction followed by feature matching, where extraction plays a critical role in determining performance. Traditional handcrafted features like SIFT (Lowe 2004) and ORB (Rublee et al. 2011) detect corners or blobs assumed to be repeatable across views, but often fail on dynamic or low-texture regions due to limited perception. Deep learning methods address these limitations by learning semantically meaningful and robust features. A key milestone is D2-Net, which introduces the joint detection-and-description framework that forms the foundation of our work (Dusmanu et al. 2019). Building on this, R2D2 emphasizes important properties like repeatability, reliability, and local saliency, which are popularly used the following methods (Revaud et al. 2019). Recent improvements include ALIKED (Zhao et al. 2022, 2023; Edstedt et al. 2024a) for better geometric consistency, MTLDesc (Wang et al. 2022) for long-range context modeling with Transformers (Vaswani et al. 2017), and XFeat (Potje et al. 2024), which decouples detection and description for speed and flexibility.

While description benefits directly from data and architecture, detection remains more challenging, as it lacks precise supervision. Instead, it relies heavily on loss design and data accuracy, making it more sensitive to noisy or weakly aligned labels (Zhang et al. 2025b).

Multi-Modal Image Matching

Unlike visible images, cross-modal images exhibit significant nonlinear appearance differences across modalities (Jiang et al. 2021). Traditional handcrafted methods (Li et al. 2022b; Hou, Liu, and Zhang 2024) address this by extracting response maps that capture common, reliable structures, then applying classic visible-image techniques for feature detection and description. For example, RIFT (Li, Hu, and Ai 2019) detects FAST (Rosten and Drummond 2006) features on phase congruency maps and constructs modality-invariant descriptors using frequency-domain features.

Most deep learning approaches (Cui et al. 2022; Zhu et al. 2023; Ye et al. 2024; Zhang et al. 2025a) adapt visible-image methods to multi-modal data. ReDFeat (Deng and Ma 2022), built upon R2D2, introduces a recoupling learning scheme to enhance feature robustness and keypoint repeatability. MIFNet (Liu et al. 2025) improves descriptors by leveraging pretrained features from Stable Diffusion, despite training solely on visible images.

However, these deep learning methods still face challenges due to the limited quantity and imprecise alignment of multi-modal datasets. Some approaches (Cui et al. 2022; Tuzcuoğlu et al. 2024) attempt to mitigate this by loading models pretrained on large-scale visible image datasets, but

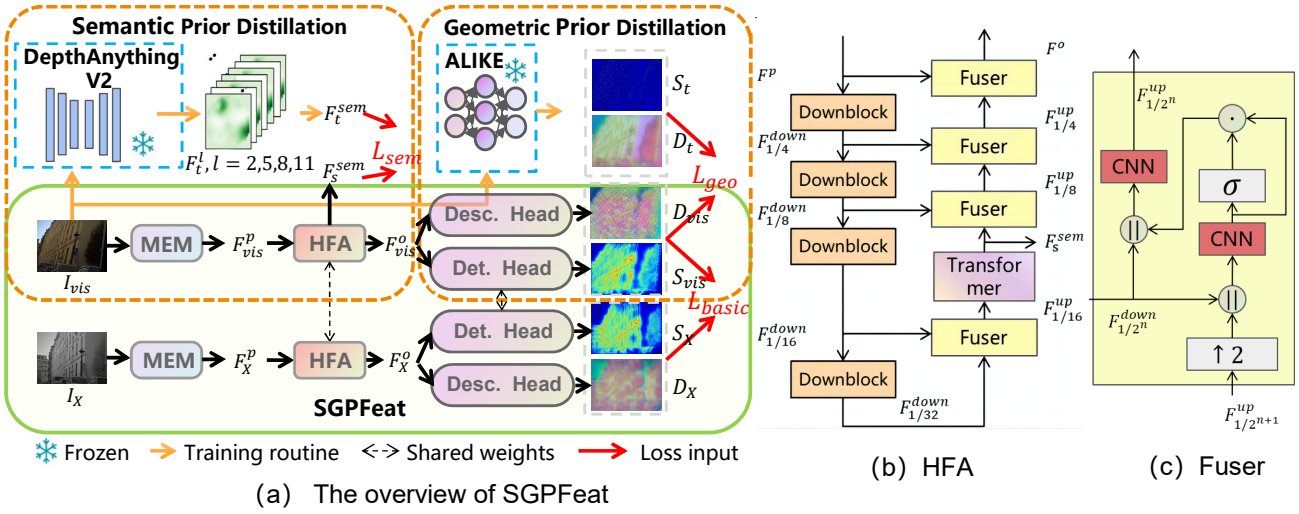

Figure 2: The framework of SGPFeat and its components.

this knowledge often diminishes after fine-tuning. XoFTR employs Masked Image Modeling (MIM) (Xie et al. 2022) self-supervision to enhance robustness without requiring extra data. Additionally, pseudo-multimodal datasets like MD-syn (Li and Snavely 2018) and data engines such as MatchAnything (He et al. 2025) improve cross-modal generalization by providing more diverse training samples. While self-supervision or full supervision on synthetic data boosts robustness, they frequently suffer from distribution shifts relative to real-world data. Therefore, this paper investigates an alternative approach to overcome the scarcity and quality issues in multi-modal training data.

Method

Overview

As shown in Fig. 2 (a), given a pair of cross-modal images $I_{vis} \in \mathbb{R}^{H\times W\times 3}$ and $I_X \in \mathbb{R}^{H\times W\times C_{in}}$, our goal is to employ a neural network to extract a keypoint detection score map $S \in \mathbb{R}^{H\times W}$ and dense descriptors $D \in \mathbb{R}^{\frac{H}{2}\times\frac{W}{2}\times 128}$ for each image, which are then used to sample sparse features for image matching. For basic feature learning, we adapt the loss functions $\mathcal{L}_{basic}$ from ReDFeat (Deng and Ma 2022). To address the scarcity and misalignment issues in training datasets, we introduce auxiliary learning tasks based on semantic and geometric prior distillation. Specifically, semantic priors from the vision foundation model DepthAnything V2 (Yang et al. 2024) are distilled into our feature extractor through feature alignment using $\mathcal{L}_{sem}$. In parallel, geometric priors from the ALIKE model (Zhao et al. 2022) are distilled using both label-level and feature-level knowledge distillation (Gou et al. 2021) by $\mathcal{L}_{geo}$.

Model Design

As shown in Fig. 2(a), we design a neural architecture with several novel modules to capture shared representations across multi-modal images and better utilize prior knowledge from teacher networks. It begins with lightweight and unshared Modality-aware Encoding Module (MEM) to mitigate appearance discrepancies across modalities. The extracted features F^p are then enhanced by Heterogeneous Feature Aggregation (HFA) modules, which expand the receptive field and integrate multi-scale context to strengthen representational capability. Finally, given enhanced features F^o a descriptor head produces dense descriptors D, while a detection head outputs the keypoint score map S, both supervised by basic losses. Key components of this design are described below.

Modality-Aware Encoding Module (MEM) Multi-modal images often exhibit sparse textures and sensor-specific noise, which reduces their structural similarity to visible-spectrum images and complicates modality-invariant feature extraction. To alleviate this modality gap, we adopt an MEM. Given an input image, MEM first applies instance normalization to mitigate the domain shift, and then feeds the normalized features into a downsampling block (Downblock) consisting of four convolutional layers, where the second layer uses a stride of 2. This design progressively extracts modality-invariant features while reducing the spatial resolution.

Heterogeneous Feature Aggregation (HFA) Module Building upon MEM, further mitigating modality variations requires receptive fields that can simultaneously capture fine-grained local details and long-range dependencies. To this end, we adopt an HFA module that combines CNNs and Transformer, enabling precise local feature encoding as well as global context modeling. Moreover, this heterogeneous design aligns with the architectures of the teacher networks, DepthAnything V2 (Transformer-based) and ALIKE (CNN-based), facilitating cross-architecture knowledge transfer and thereby improving distillation performance.

Specifically, HFA module is built upon a U-Net architecture (Li, Zhang, and Ma 2024), which efficiently provides a large receptive field, as shown in Fig. 2(b). The encoder branch extracts a feature pyramid from $1/2$ to $1/32$ resolu-

tion using a series of Downblocks, which are similar to those used in the MEM, while Transformer for long-range context modeling is applied at the $1/16$ scale to balance modeling capacity and computational cost.

To adaptively fuse multi-scale CNN and Transformer features, we introduce an Adaptive Fuser as shown in Fig. 2(c). Given a lower-resolution feature $F^{up}_{1/2^{(n+1)}}$ from the decoder branch and a higher-resolution feature $F^{down}_{1/2^n}$ from the encoder branch, the fuser upsamples $F^{up}_{1/2^{(n+1)}}$ via bilinear interpolation, concatenates it with $F^{down}_{1/2^n}$, computes importance scores for information selection, and reweights the features accordingly. Finally, the fused representation is compressed to the target dimension, with a shortcut connection preserving the original concatenated input:

$$\begin{aligned} F_{\text{cat}} &= F^{up}_{1/2^{(n+1)}} || F^{down}_{1/2^n}, \\ G &= \sigma\big(f_g(F_{\text{cat}})\big), \\ F^{up}_{1/2^n} &= f_c\big((G \odot F_{\text{cat}}) || F_{\text{cat}}\big), \end{aligned} \tag{1}$$

where σ denotes the sigmoid activation, $\odot$ denotes the element-wise multiplication, f denotes a light-weight CNNs, $||$ denotes concatenation.

Basic Constraints To implement basic feature learning, we adopt the loss function of ReDFeat, denoted as $\mathcal{L}_{basic}$. ReDFeat introduces a carefully designed recoupling strategy that jointly optimizes detection and description losses, enabling their mutual reinforcement. To further improve the stability of descriptor training, we replace the original angular descriptor loss in $\mathcal{L}_{basic}$ with a cosine similarity loss. Please refer to ReDFeat for more details.

Semantic Prior Distillation

Motivation The performance of deep learning models is highly dependent on the quality and quantity of training data. In multi-modal image matching, however, progress is often hindered by the scarcity and limited quality of available datasets—an issue particularly severe for Transformer-based architectures (Dosovitskiy et al. 2020), which are inherently data-hungry. Our method faces the same limitation. To address this challenge, we propose to transfer rich semantic knowledge from vision foundation models trained on large-scale and diverse datasets. Specifically, we leverage DepthAnything V2 (Yang et al. 2024), a state-of-the-art vision foundation model derived from the powerful DINO V2 teacher (Caron et al. 2021) and large-scale data.

As shown in Fig. 3, we visualize the intermediate features of DepthAnything V2 by reducing the 384-dimensional feature channels to RGB space using Principal Component Analysis (PCA). These features demonstrate strong consistency for objects of the same category, even across diverse scenes and depth variations. We regard this property as a semantic knowledge, as it captures object-centric regularities that remain invariant under complex scene changes. Incorporating such a prior facilitates our model's understanding of cross-modal images, enabling it to capture shared patterns and textures despite significant appearance differences across modalities. To exploit this prior, we extract and fuse

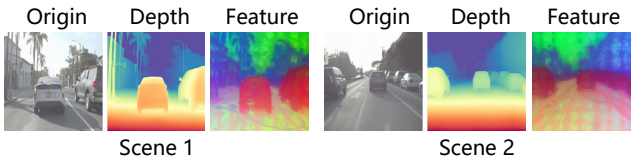

Figure 3: Visualization of DepthAnything V2 depth map and intermediate layer features.

multi-layer intermediate features from the teacher model and align them with the Transformer layer of SGPFeat.

Semantic Feature Distillation To effectively leverage the semantic prior in DepthAnything V2, we extract multi-stage intermediate features from its lightweight variant, DepthAnything V2-S. These features are fused along the channel dimension to form a unified representation, which is then aligned with the student's semantic feature map F^{sem}:

$$F^{\text{sem}}_t = \textstyle\sum_{l\in\{2,5,8,11\}} w_l \cdot F^l_t, \tag{2}$$

where F^l_t is the feature from the l-th Transformer block of the teacher, and w is a learnable weight. The fused teacher feature F^{sem}_t is projected to match the channel dimension of the student's semantic output F^{sem} from the HFA module. Semantic distillation is then performed by maximizing the similarity between teacher and student features:

$$\mathcal{L}_{\text{sem}} = 1 - \frac{1}{H'W'} \sum_{i=1}^{H'} \sum_{j=1}^{W'} \frac{(F^{i,j}_s)^\top F^{i,j}_t}{\|F^{i,j}_s\| \|F^{i,j}_t\|}, \tag{3}$$

where $H' = H/16$, and $W' = W/16$. To further transfer teancher's semantic knowledge across modalities, we rely on shared weights and the alignment constraints imposed by the basic loss $\mathcal{L}_{basic}$. Such distillation allows the student network—our feature extractor—to absorb semantic priors from large-scale data and generalize them to other modalities and capture common structural patterns.

Geometric Prior Distillation

Motivation In most multi-modal datasets (Deng and Ma 2022), image pairs are typically assumed to follow a homography transformation. However, this assumption holds true only in limited scenarios, such as satellite imagery. As illustrated in Fig. 4(a), applying a global homography to align cross-modal images often leads to local misalignments, particularly around depth discontinuities, which are common locations for keypoints. These geometric inaccuracies undermine the effectiveness of feature learning, which relies on accurate geometric correspondence. In contrast, visible-light image datasets are often constructed through 3D reconstruction and reprojection pipelines (Li and Snavely 2018), resulting in more geocentrically accurate alignment that better supports geometry-based feature learning.

ALIKE is among the pioneering methods for accurate keypoint detection, introducing Differentiable Keypoint Detection (DKD) and minimizing reprojection error during training. Leveraging large-scale datasets with precise geometric annotations, ALIKE effectively learns rich geocentric knowledge that enables reliable keypoint detection and

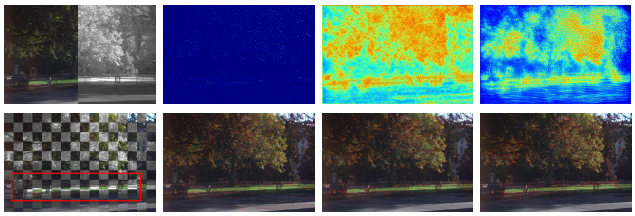

Figure 4: Visualization of misalignment in the groudtruth and comparison of keypoint detection.

robust feature description. To mitigate the geometric inaccuracies inherent in multi-modal data, we propose to distill these geometric priors from ALIKE into our model, particularly the detection score via label distillation. This process aligns the distribution of detection confidences, thereby transferring precise keypoint localization knowledge to improve performance on noisy multi-modal inputs.

Geometric Label Distillation While globally normalizing detection score maps and applying Kullback-Leibler (KL) divergence between teacher and student outputs is conceptually simple, it often causes oversmoothing. Because the normalization denominator aggregates values across the entire map, the resulting distributions become excessively flattened, diminishing contrast and keypoint-specific signals, which may ultimately lead to distillation collapse. To address this, we partition the score S into a set of overlapping local patches $\{p^i, i = 0, 1, ..., L\}$ using an unfold operation with a kernel size of 5×5 and a stride of 2. Each patch is then normalized independently using Softmax:

$$\begin{aligned} \hat{p}_t^i &= \mathrm{Softmax}(\tau \cdot \sigma^{-1}(p_t^i)), \\ \hat{p}_s^i &= \mathrm{Softmax}(\tau \cdot \sigma^{-1}(p_s^i)), \end{aligned} \tag{4}$$

where $\tau = 5$ is a temperature parameter that controls the sharpness of the resulting distribution. Finally, the final score map distillation loss is:

$$\mathcal{L}_{\text{geo-det}} = 1/L \sum_i w_i \cdot D_{\mathrm{KL}} \left(\hat{p}_t^i \parallel \hat{p}_s^i \right), \tag{5}$$

where each weight w_i is computed as the sum of the values within p^i. This weighting scheme encourages the optimization to focus on salient regions in the visible images, effectively acting as an attention mechanism. The proposed formulation preserves the sparsity and local characteristics of keypoints extracted from the score map while ensuring numerical stability during cross-modal training. As shown in Fig. 4, such distillation enables our network to detect more reliable and accurate keypoints along corners and edges, thereby improving image matching performance.

Geometric Feature Distillation The features of ALIKE also contain rich geometric knowledge. To improve robustness of the feature to geometric variants, we additionally maximize the similarity between teacher descriptor map D_t from ALIKE and student descriptor map D_s from SGPFeat:

$$\mathcal{L}_{\text{geo-desc}} = 1 - \frac{1}{HW} \sum_{i=1}^{H} \sum_{j=1}^{W} \frac{(D_s^{i,j})^\top D_t^{i,j}}{\|D_s^{i,j}\| \|D_t^{i,j}\|}. \tag{6}$$

Note that, the resolution of D_s is restored to the original image resolution using bilinear upsampling to align it with D_t. Since ALIKE is trained exclusively on visible images with a lightweight backbone, it lacks robustness when directly serving as a teacher for other modalities. The geometric distillation is applied exclusively to the visible image branch. Finally, the overall training loss function is defined as:

$$\begin{aligned} \mathcal{L}_{\text{total}} &= \mathcal{L}_{\text{basic}} + \mathcal{L}_{\text{sem}} + \mathcal{L}_{\text{geo}}, \\ \mathcal{L}_{\text{geo}} &= \mathcal{L}_{\text{geo-det}} + \mathcal{L}_{\text{geo-desc}}. \end{aligned} \tag{7}$$

Experiments

Settings

Our SGPFeat is trained on the VIS-NIR and VIS-IR subsets of the Multi-modal Feature Evaluation benchmark (Deng and Ma 2022). The VIS-NIR subset contains only 345 pairs of visible and near-infrared (NIR) images, while the VIS-IR subset includes 211 pairs of visible and infrared (IR) images. These datasets are significantly smaller and noisier than those used to train the teacher models, DepthAnything V2-S (Yang et al. 2024) and ALIKE-L (Zhao et al. 2022). Training image pairs are first resized so that their shortest side is 640 pixels, followed by random homography transforms and cropping to 448×448. For the VIS-IR setting, both SemD and GeoD are applied only to the visible branch. For the VIS-NIR setting, leveraging the robustness of DepthAnything V2 to modalities analogous to visible images, we further extend SemD to the NIR branch. Models are trained with a batch size of 2 using the AdamW optimizer (Loshchilov and Hutter 2017), a learning rate of 0.0001, and a weight decay of 0.01. The learning rate follows a cosine annealing schedule over 800 epochs, with a minimum value of 1×10^{-7}.

We first evaluate the overall effectiveness of SGPFeat through homography estimation on the Multi-modal Feature Evaluation and relative camera pose estimation on METU_VisTIR (Tuzcuoğlu et al. 2024). To test its zero-shot capability gained through knowledge distillation, we test SGPFeat on unseen real-world medical datasets, including CT-MRI (Gu et al. 2024) and EMA-OCTA (Wang et al. 2024a). For the above tasks, SGPFeat is compared with the established methods below, including sparse feature-based approaches RIFT (Li, Hu, and Ai 2019), ALIKE (Zhao et al. 2022), ReDFeat (Deng and Ma 2022), POS-GIFT (Hou, Liu, and Zhang 2024), MIFNet (Liu et al. 2025), and the semi-dense method XoFTR (Tuzcuoğlu et al. 2024). All experiments are conducted on a single NVIDIA RTX 3090 GPU.

Homography Estimation

The Multi-modal Feature Evaluation benchmark (Deng and Ma 2022), proposed by ReDFeat (Deng and Ma 2022), focuses on estimating homography transforms between multi-modal images. These images are manually aligned and then warped using synthetic transformations. For homography estimation, 4096 keypoints are extracted by sparse feature extraction methods. With the exception of MINIMA, which employs a graph network for feature matching, the

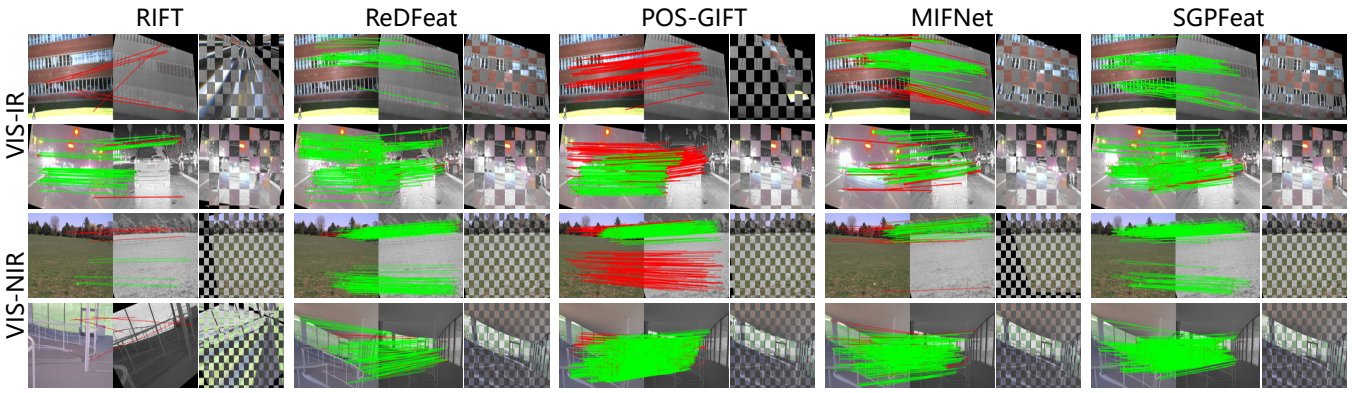

Figure 5: Visualization of matching and homography estimation. The top 1024 keypoints are extracted for sparse feature extraction algorithms. Correct matches with R_E <10px are drawn in green, while the others are drawn in red.

Method	VIS-IR			VIS-NIR		
	SRR (↑%)	R_{avg} (↓px)	CLR (↓%)	SRR (↑%)	R_{avg} (↓px)	CLR (↓%)
RIFT	66.0	5.48	6.4	25.0	5.14	37.9
ReDFeat	**97.8**	**2.68**	**0**	*62.9*	*4.53*	0.8
POS-GIFT	55.3	5.27	4.3	3.0	4.75	31.8
MIFNet	91.5	3.44	**0**	61.4	4.96	*0.7*
SGPFeat	*95.7*	*3.23*	**0**	**81.1**	**3.01**	**0**
MINIMA	95.7	2.31	0	87.1	2.71	0
XoFTR	93.6	2.65	0	90.9	2.43	0

Table 1: Multi-modal Homography Estimation. MINIMA and XoFTR are shown in a separate column due to their use of additional feature matching networks. The best and second-best scores are highlighted in bold and italic.

other sparse methods perform feature matching using Mutual Nearest Neighbor (MNN) matching. For semi-dense matching, we follow the official XoFTR pipeline. Given the resulting matches, RANSAC (Fischler and Bolles 1981) is applied to estimate the homography matrix.

The estimation accuracy is assessed using the reprojection error R_E, computed from the four image corners (VIS-NIR) or from annotated landmarks (VIS-IR). Based on R_E, we report three metrics: the Successful Registration Rate (SRR), *i.e.*, the proportion of images with $R_E < 10$ pixels; the Collapsed Rate (CLR), *i.e.*, the proportion with $R_E > 100$ pixels; and the average reprojection error R_{avg}, computed over images with $R_E < 10$ pixels. As summarized in Tab. 1, SGPFeat surpasses methods based on MNN across all metrics, underscoring the effectiveness of our approach. Notably, the substantial improvements in SRR and R_{avg} indicate that our method delivers both high robustness and high accuracy.

Some qualitative results are presented in Fig. 5. As shown, our method extracts more reliable and accurately localized keypoints that are evenly distributed across the images. Moreover, benefiting from knowledge distillation, our descriptors exhibit improved robustness, resulting in a higher number of inlier matches.

Method	Cloudy_Cloudy (↑%)			Cloudy_Sunny (↑%)		
	@5°	@10°	@20°	@5°	@10°	@20°
RIFT	0.15	0.94	1.98	0.00	0.0	0.67
ReDFeat	0.82	2.82	8.14	*0.73*	*2.34*	5.81
MIFNet	*1.11*	*3.36*	*9.61*	0.48	2.15	*7.80*
ALIKE	0.00	0.00	0.44	0.00	0.06	0.44
SGPFeat	**2.55**	**8.09**	**19.31**	**0.82**	**3.62**	**10.74**
XoFTR	22.17	38.76	54.95	13.21	28.05	45.05

Table 2: Relative Pose Estimation on METU_VisTIR. Areas Under Curve (AUCs) at 5°, 10°, 20° are reported.

Relative Pose Estimation

The METU_VisTIR dataset, introduced by XoFTR, comprises visible and infrared image pairs acquired under diverse viewpoints and environmental conditions (Tuzcuoğlu et al. 2024). Due to the significant viewpoint variations, homography-based approximations are unreliable, necessitating relative camera pose estimation, which in turn demands robust and accurate geometric representations. Following the aforementioned feature extraction and matching pipeline, we estimate the essential matrix using RANSAC after calibrating with known intrinsic parameters. Relative camera poses are subsequently recovered from the essential matrix, and the estimation accuracy is quantified using the Area Under the Curve (AUC) at various thresholds. As shown in Tab. 2, SGPFeat achieves notably superior results compared to existing methods, demonstrating its ability to effectively transfer geometric priors from unimodal to multimodal tasks through distillation using only a small amount of VIS-IR homography data.

Zero-Shot Analysis

The EMA-OCTA dataset from MEMO, is a multimodal retinal image benchmark designed to address large differences in vascular density and non-rigid deformations between imaging modalities. We evaluate homography estimation on this dataset using the Successful Registration Rate (SRR) and the average reprojection error R_{avg}, both computed from manually annotated landmarks at the scaled res-

Method (Train data)	CT-MRI		EMA-OCTA	
	SRR (↑%)	R_{avg} (↓px)	SRR (↑%)	R_{avg} (↓px)
POS-GIFT	62.5	7.42	36.7	3.48
MIFNet (Retina)	**79.2**	**6.73**	36.7	**3.35**
MIFNet (Remote)	45.8	7.80	*46.7*	*3.40*
SGPFeat (VIS-IR)	**79.2**	*6.88*	**50.0**	4.50
SGPFeat (VIS-NIR)	58.3	7.81	40.0	3.58

Table 3: Zero-shot performance on medical images.

Method	SemD	GeoD		VIS-NIR		
		det.	desc.	SRR (↑%)	R_{avg} (↓px)	CLR (↓%)
(a)				34.1	6.10	10.6
(b)	✓			45.5	5.41	0.8
(c)		✓	✓	72.7	4.12	0
(d)	✓	✓		71.2	3.70	1.5
(e)	✓	✓	✓	81.1	3.65	0

Table 4: Ablation study for different distillation losses.

Figure 6: Comparison of intermediate features of DepthAnything (DepAny) V2 and DINO V2.

olution defined in MEMO. For the CT-MRI dataset, registration accuracy is similarly assessed using the mean reprojection error of four image corners.

As shown in Tab. 3, our model trained on VIS-IR achieves performance comparable to MIFNet (Retina), which leverages diffusion-based distribution alignment on the training split of this dataset (Liu et al. 2025). As the SRR increases, since R_{avg} is computed based on SR images, the challenging non-rigid deformation cases in EMA-OCTA result in a higher overall R_{avg}. Furthermore, our model outperforms MIFNet (Remote), trained on remote sensing imagery. Since IR shares structural and energy-based modality traits with CT-MRI and EMA-OCTA, while NIR provides richer textures, the VIS-IR variant generalizes better.

Discussion

Ablation Study

We conduct ablation studies to rigorously assess the effectiveness of the two core components of our framework, namely semantic and geometric distillation. Experiments are performed on the VIS-NIR dataset using 1024 keypoints. The SGPFeat model trained solely with the basic constraints is adopted as the baseline. As summarized in Tab. 4, incorporating either semantic or geometric distillation individually yields notable performance improvements, validating the contribution of each component. Furthermore, combining both distillation strategies achieves the highest overall performance, demonstrating their complementary nature. Qualitative analyses additionally reveal that semantic distillation improves the cross-modal consistency of feature descriptors, whereas geometric distillation refines the spatial distribution and repeatability of detected keypoints by guiding them toward structurally stable and reliable regions.

Method	SemD	GeoD	VIS-IR		
			SRR (↑%)	R_{avg} (↓px)	CLR (↓%)
(f)	DepAny V2	ALIKED	87.2	4.23	0
(g)	DepAny V2	RDD	44.7	4.87	17.0
(h)	DepAny V2	ALIKE	87.2	3.94	0

Table 5: Replacement study of geometric teacher networks.

Teacher Network Replacement

DINO V2 (Oquab et al. 2023), a widely used vision foundation model and the teacher of DepthAnything (DepAny) V2, was considered as an alternative semantic teacher. However, distillation from DINO V2 produces less competitive results. Visualization of intermediate features in Fig. 6 shows that DepthAnything V2 captures smoother and more coherent responses, whereas DINO V2 exhibits pronounced discrete noise in certain regions, such as the sky. Since our task requires clean, stable semantic correspondences, DepthAnything V2 is a more suitable semantic prior.

Moreover, multiple alternatives exist for geometric distillation, such as the state-of-the-art feature extraction method, RDD (Chen et al. 2025), and the popular lightweight approach, ALIKED (Zhao et al. 2023). We replaced the ALIKE teacher with these methods, and the results in Tab. 5 show that ALIKE achieves the best performance. A possible reason is that RDD employs a deformable local Transformer, which introduces a large architectural gap compared to our model, while ALIKED, being lightweight, provides limited geometric knowledge for effective transfer.

Conclusion

We propose a brand new method called SGPFeat, which introduces semantic priors and geometric priors from unimodal models into cross-modal training under the constraints of dedicated loss functions. Addressing the challenge of the scarcity and the inaccurate alignment of multimodal datasets, SGPFeat extracts features with shared semantic structures across different modalities and detects accurate correspondences on noisy aligned multi-modal pairs. A series of experiments demonstrate that SGPFeat achieves excellent performance in homography estimation and relative pose estimation. Additionally, supplementary experiments verify the compatibility of the selection of teacher networks and the design of distillation losses.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (62506268, 62276192), and the Natural Science Foundation of Jiangsu Province (BK20250454).

References

Caron, M.; Touvron, H.; Misra, I.; Jégou, H.; Mairal, J.; Bojanowski, P.; and Joulin, A. 2021. Emerging Properties in Self-Supervised Vision Transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

Chen, G.; Fu, T.; Chen, H.; Teng, W.; Xiao, H.; and Zhao, Y. 2025. RDD: Robust Feature Detector and Descriptor using Deformable Transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6394–6403.

Cui, S.; Ma, A.; Wan, Y.; Zhong, Y.; Luo, B.; and Xu, M. 2022. Cross-Modality Image Matching Network With Modality-Invariant Feature Representation for Airborne-Ground Thermal Infrared and Visible Datasets. *IEEE Transactions on Geoscience and Remote Sensing*, 60: 5606414.

Deng, Y.; and Ma, J. 2022. ReDFeat: Recoupling detection and description for multimodal feature learning. *IEEE Transactions on Image Processing*, 32: 591–602.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Dusmanu, M.; Rocco, I.; Pajdla, T.; Pollefeys, M.; Sivic, J.; Torii, A.; and Sattler, T. 2019. D2-net: A trainable cnn for joint description and detection of local features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8092–8101.

Edstedt, J.; Bökman, G.; Wadenbäck, M.; and Felsberg, M. 2024a. DeDoDe: Detect, Don't Describe—Describe, Don't Detect for Local Feature Matching. In *International Conference on 3D Vision*, 148–157.

Edstedt, J.; Sun, Q.; Bökman, G.; Wadenbäck, M.; and Felsberg, M. 2024b. RoMa: Robust dense feature matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19790–19800.

Fischler, M. A.; and Bolles, R. C. 1981. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6): 381–395.

Gou, J.; Yu, B.; Maybank, S. J.; and Tao, D. 2021. Knowledge distillation: A survey. *International Journal of Computer Vision*, 129(6): 1789–1819.

Gu, X.; Wang, L.; Deng, Z.; Cao, Y.; Huang, X.; and min Zhu, Y. 2024. Adaptive spatial and frequency experts fusion network for medical image fusion. *Biomedical Signal Processing and Control*, 96: 106478.

Guo, Z.; Li, X.; Huang, H.; Guo, N.; and Li, Q. 2019. Deep learning-based image segmentation on multimodal medical imaging. *IEEE Transactions on Radiation and Plasma Medical Sciences*, 3(2): 162–169.

He, X.; Yu, H.; Peng, S.; Tan, D.; Shen, Z.; Bao, H.; and Zhou, X. 2025. MatchAnything: Universal Cross-Modality Image Matching with Large-Scale Pre-Training. *arXiv preprint arXiv:2501.07556*.

Hou, Z.; Liu, Y.; and Zhang, L. 2024. POS-GIFT: A geometric and intensity-invariant feature transformation for multimodal images. *Information Fusion*, 102: 102027.

Jiang, X.; Ma, J.; Xiao, G.; Shao, Z.; and Guo, X. 2021. A review of multimodal image matching: Methods and applications. *Information Fusion*, 73: 22–71.

Kovesi, P. 2003. Phase congruency detects corners and edges. In *The Australian Pattern Recognition Society Conference: DICTA*, volume 2003.

Li, J.; Hong, D.; Gao, L.; Yao, J.; Zheng, K.; Zhang, B.; and Chanussot, J. 2022a. Deep learning in multimodal remote sensing data fusion: A comprehensive review. *International Journal of Applied Earth Observation and Geoinformation*, 112: 102926.

Li, J.; Hu, Q.; and Ai, M. 2019. RIFT: Multi-modal image matching based on radiation-invariant feature transform. *IEEE Transactions on Image Processing*, 29: 3296–3310.

Li, J.; Xu, W.; Shi, P.; Zhang, Y.; and Hu, Q. 2022b. LNIFT: Locally normalized image for rotation invariant multimodal feature matching. *IEEE Transactions on Geoscience and Remote Sensing*, 60: 1–14.

Li, Z.; and Snavely, N. 2018. MegaDepth: Learning Single-View Depth Prediction from Internet Photos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Li, Z.; Zhang, S.; and Ma, J. 2024. U-Match: Exploring Hierarchy-aware Local Context for Two-view Correspondence Learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12): 10960–10977.

Liu, Y.; Sun, Z.; Yu, B.; Zhao, Y.; Du, B.; Xu, Y.; and Cheng, J. 2025. MIFNet: Learning Modality-Invariant Features for Generalizable Multimodal Image Matching. *IEEE Transactions on Image Processing*, 34: 3593–3608.

Loshchilov, I.; and Hutter, F. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Lowe, D. G. 2004. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60: 91–110.

Mishchuk, A.; Mishkin, D.; Radenovic, F.; and Matas, J. 2017. Working hard to know your neighbor's margins: Local descriptor learning loss. In *Advances in Neural Information Processing Systems*, volume 30, 1–12.

Oquab, M.; Darcet, T.; Moutakanni, T.; Vo, H. V.; Szafraniec, M.; Khalidov, V.; Fernandez, P.; Haziza, D.; Massa, F.; El-Nouby, A.; Howes, R.; Huang, P.-Y.; Xu, H.; Sharma, V.; Li, S.-W.; Galuba, W.; Rabbat, M.; Assran, M.; Ballas, N.; Synnaeve, G.; Misra, I.; Jegou, H.; Mairal, J.; Labatut, P.; Joulin, A.; and Bojanowski, P. 2023. DINOv2: Learning Robust Visual Features without Supervision.

Potje, G.; Cadar, F.; Araujo, A.; Martins, R.; and Nascimento, E. R. 2024. XFeat: Accelerated Features for Lightweight Image Matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2682–2691.

Ren, J.; Jiang, X.; Li, Z.; Liang, D.; Zhou, X.; and Bai, X. 2025. Minima: Modality invariant image matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1–8.

Revaud, J.; De Souza, C.; Humenberger, M.; and Weinzaepfel, P. 2019. R2d2: Reliable and repeatable detector and descriptor. In *Advances in Neural Information Processing Systems*, volume 32, 1–11.

Rosten, E.; and Drummond, T. 2006. Machine learning for high-speed corner detection. In *European Conference on Computer Vision*, 430–443. Springer.

Rublee, E.; Rabaud, V.; Konolige, K.; and Bradski, G. 2011. ORB: An efficient alternative to SIFT or SURF. In *Proceedings of the IEEE International Conference on Computer Vision*, 2564–2571.

Sarlin, P.-E.; Cadena, C.; Siegwart, R.; and Dymczyk, M. 2019. From coarse to fine: Robust hierarchical localization at large scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12716–12725.

Sarlin, P.-E.; DeTone, D.; Malisiewicz, T.; and Rabinovich, A. 2020. Superglue: Learning feature matching with graph neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4938–4947.

Schonberger, J. L.; and Frahm, J.-M. 2016. Structure-from-motion revisited. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4104–4113.

Sun, J.; Shen, Z.; Wang, Y.; Bao, H.; and Zhou, X. 2021. LoFTR: Detector-free local feature matching with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8922–8931.

Tuzcuoğlu, Ö.; Köksal, A.; Sofu, B.; Kalkan, S.; and Alatan, A. A. 2024. Xoftr: Cross-modal feature matching transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 4275–4286.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 1–11.

Wang, C.; Xu, R.; Zhang, Y.; Xu, S.; Meng, W.; Fan, B.; and Zhang, X. 2022. MTLDesc: Looking wider to describe better. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 2388–2396.

Wang, C.-Y.; Sadrieh, F. K.; Shen, Y.-T.; Chen, S.-E.; Kim, S.; Chen, V.; Raghavendra, A.; Wang, D.; Saeedi, O.; and Tao, Y. 2024a. MEMO: dataset and methods for robust multimodal retinal image registration with large or small vessel density differences. *Biomedical Optics Express*, 15(5): 3457–3479.

Wang, Y.; He, X.; Peng, S.; Tan, D.; and Zhou, X. 2024b. 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.

Xie, Z.; Zhang, Z.; Cao, Y.; Lin, Y.; Bao, J.; Yao, Z.; Dai, Q.; and Hu, H. 2022. Simmim: A simple framework for masked image modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9653–9663.

Yang, L.; Kang, B.; Huang, Z.; Zhao, Z.; Xu, X.; Feng, J.; and Zhao, H. 2024. Depth anything v2. *Advances in Neural Information Processing Systems*, 37: 21875–21911.

Ye, Y.; Yang, C.; Gong, G.; Yang, P.; Quan, D.; and Li, J. 2024. Robust optical and SAR image matching using attention-enhanced structural features. *IEEE Transactions on Geoscience and Remote Sensing*, 62: 1–12.

Zhang, K.; Deng, Y.; Ma, J.; and Favaro, P. 2025a. Adapting dense matching for homography estimation with grid-based acceleration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6294–6303.

Zhang, S.; Li, Z.; Zhang, K.; Lu, Y.; Deng, Y.; Tang, L.; Jiang, X.; and Ma, J. 2025b. Deep Learning Reforms Image Matching: A Survey and Outlook. *arXiv preprint arXiv:2506.04619*.

Zhao, X.; Wu, X.; Chen, W.; Chen, P. C.; Xu, Q.; and Li, Z. 2023. Aliked: A lighter keypoint and descriptor extraction network via deformable transformation. *IEEE Transactions on Instrumentation and Measurement*, 72: 1–16.

Zhao, X.; Wu, X.; Miao, J.; Chen, W.; Chen, P. C.; and Li, Z. 2022. Alike: Accurate and lightweight keypoint detection and descriptor extraction. *IEEE Transactions on Multimedia*, 25: 3101–3112.

Zhu, B.; Yang, C.; Dai, J.; Fan, J.; Qin, Y.; and Ye, Y. 2023. R_2FD_2: fast and robust matching of multimodal remote sensing images via repeatable feature detector and rotation-invariant feature descriptor. *IEEE Transactions on Geoscience and Remote Sensing*, 61: 1–15.