

SemFE:A Feature Matching Method for Learnable Local Semantic Feature Enhancement in Multimodal images

Rongrui Teng

Yunnan University

Yun Liao

LiaoYun@ynu.edu.cn

Yunnan University

Wei Wang

Yunnan University

Qing Duan

Yunnan University

Junhui Liu

Yunnan University

Fangwei Jin

Yunnan University

Yunpeng Li

Yunnan University

Xu Qian

Yunnan University

Article

Keywords: Feature Matching, Multi-modal Image, Semantic, Transformer, Attention

Posted Date: November 13th, 2024

DOI: <https://doi.org/10.21203/rs.3.rs-5275064/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

SemFE: A Feature Matching Method for Learnable Local Semantic Feature Enhancement in Multimodal Images

Rongrui Teng^{1,3+}, Yun Liao^{1,2,3+✉}, Wei Wang^{1,3}, Qing Duan^{1,3}, Junhui Liu^{1,3}, Fangwei Jin¹, Yunpeng Li¹, and Xu Qian¹

¹Yunnan University, Kunming, 650506, China.

²Yunnan Lanyi Network Technology Co, Kunming, 650000, China.

³Yunnan Key Laboratory of Software Engineering, Kunming, 650506, China.

⁺These authors contributed equally to this work

✉email:LiaoYun@ynu.edu.cn

ABSTRACT

Multimodal image matching has been a critical challenge within the field of computer vision over an extended period. In recent years, detector free methods have received widespread attention for achieving high matching accuracy. However, these methods often fail to fully exploit the semantic information within images, which is crucial for achieving accurate matching. To overcome this limitation, we propose SemFE, a multimodal image matching framework that leverages learnable local semantic feature enhancement. In SemFE, we designed a dynamic semantic feature extraction module to capture semantic information, along with a semantic information enhancement module to refine semantically related features. Additionally, we developed a multi-scale feature fusion backbone integrated with a transformer to further enhance feature extraction. Comprehensive experimental results demonstrate that SemFE consistently outperforms competing methods across various multimodal image datasets.

Keywords: Feature Matching, Multi-modal Image, Semantic, Transformer, Attention

Introduction

Feature matching is increasingly becoming a key task in computer vision, referring to the process of determining similarity or correspondence between images or image regions by comparing local features (such as key points or descriptors). Typically, in feature matching, significant structures or local areas in the image—such as corners, edges, or textures—are first extracted. Then, the descriptors or feature vectors of these features are compared to determine the degree of similarity or correspondence between them. Multimodal image matching extends this task to images captured by different sensors or under varying physical conditions, where the goal is to find corresponding pixels between different modalities, such as visible light images, infrared images, or lidar scans. Current mainstream feature matching approaches can generally be classified into two categories: detector-based methods^{1–12} and detector-free methods^{13–28}.

The detector-based approach consists of several key steps. First, keypoint detection identifies salient locations in the image, which are critical for subsequent matching. Subsequently, feature extraction converts the local area surrounding each keypoint into a feature vector using methods like CNNs²⁹. These feature vectors are then compared using similarity metrics like Euclidean distance³⁰, Hamming distance³¹, or cosine similarity³². SuperGlue¹² improves this process by leveraging graph neural networks and attention mechanisms to optimize feature matching between image pairs. Finally, match verification is conducted through methods like nearest neighbor search³³ or the ratio test³⁴, followed by consistency checks to filter out unreliable matches and enhance overall accuracy.

However, detector-based methods face challenges when handling low-quality images, complex backgrounds, or images with significant noise. Since they primarily focus on local features, their performance may degrade in challenging scenarios, such as when images have small overlapping areas or large textureless regions.

Detector-free methods, which have gained considerable attention in recent years, bypass explicit keypoint detection and instead employ end-to-end network architectures to process dense feature maps directly. Notable approaches include semi-dense matching techniques such as LOFTR¹⁹, TopicFM²⁰, TopicFM+²¹, Patch2Pix²², AdaMatcher²³, ASpanFormer²⁴, and MatchFormer²⁵, as well as fully dense matching methods like DKM²⁶, Roma²⁷, and PDCnet²⁸. These methods generally

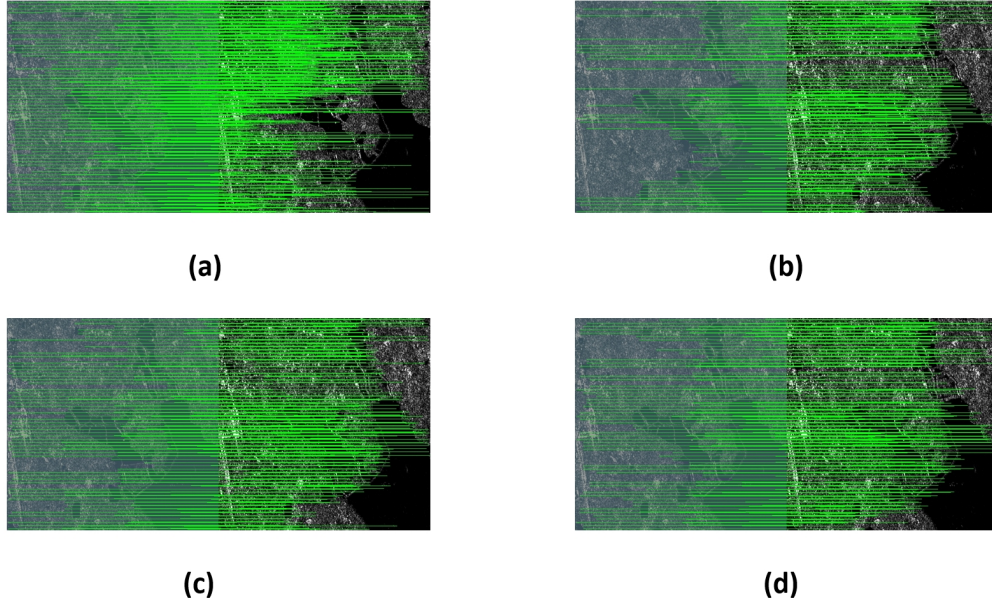


Figure 1. Comparison of SemFE with other excellent methods. (From top left to bottom right: SemFE (a), LoFTR (b)¹⁹, FeMIP (c)³⁵, and FmCFA (d).) SemFE is better at finding the correspondence between multimodal image data.

follow three key steps: First, feature extraction is performed across the entire image without relying on keypoint detection. Next, coarse matching is carried out by calculating feature similarity on low-resolution feature maps, which establishes initial matching pairs. Finally, fine matching refines these results on high-resolution feature maps, improving accuracy and robustness through geometric and cycle consistency checks. However, most of these methods tend to overlook the integration of semantic information. Semantic information is essential for numerous tasks within the domain of computer vision. For instance, in tasks such as image classification, image generation, semantic segmentation, and object detection, leveraging semantic information significantly enhances model performance. We argue that semantic information is equally vital in image matching tasks, particularly in multimodal image matching, where texture-based features may be insufficient. Currently, the two primary challenges related to semantic information are how to effectively extract and utilize it.

To tackle the first challenge, we designed a semantic feature extraction (SFE) module that dynamically captures semantic information. This module ensures the extraction of higher-quality semantic features. To tackle the second challenge, we introduced a novel semantic information enhancement (SIE) module, specifically designed to emphasize semantically important regions within the image. This module ensures that regions with significant contextual meaning are more prominently represented during the final matching process.

A significant number of experiments demonstrate that our SemFE method achieves impressive results in multimodal image matching tasks. Particularly when incorporating semantic information, the semantic information enhancement module enables more accurate analysis and comparison of details across different semantics in the images, yielding more robust matching results. Our partial matching results are presented in Fig 1. This breakthrough offers a novel solution for feature matching in multimodal scenarios and further underscores the critical role of semantic information in enhancing the performance of computer vision tasks. Additionally, Fig 2 illustrates the results of the semantic information assignment.

The key contributions of this paper are outlined below:

1. Multi-scale feature extraction (MFE) backbone: We designed a multi-scale feature extraction backbone that incorporates the attention mechanism, enabling precise extraction and analysis of image features at various scales, thereby enhancing matching accuracy.
2. Semantic feature extraction (SFE) module: We designed a semantic feature extraction module that can dynamically extract semantic information according to the image size to ensure that semantic information can be effectively captured at different scales.
3. Semantic information enhancement (SIE) module: We developed a novel semantic information enhancement module that amplifies features related to semantic information, making regions with crucial semantic content more prominent during the matching process.
4. Effectiveness verification in multimodal scenarios: Extensive experiments validate the efficacy of our proposed modules in multimodal image matching tasks, resulting in a substantial improvement in matching performance.

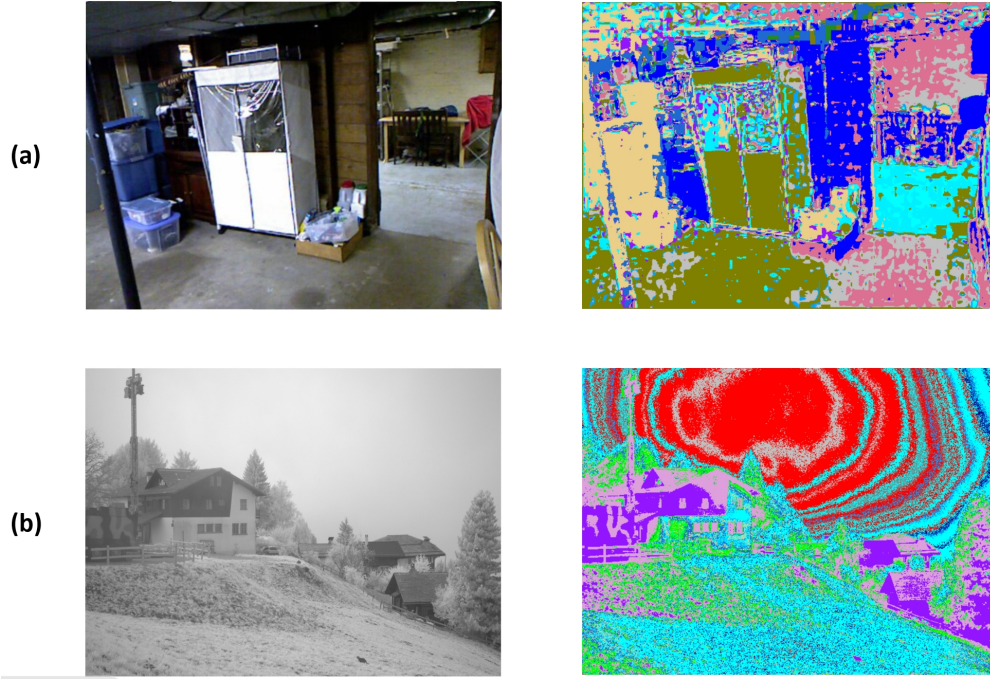


Figure 2. Original Image and Semantic Information Assignment Results.(Datasets: NYU-Depth V2 (a), RGB-NIR Scene (b))

Related Work

Detector-Base method

Detector-based approaches are essential in the realm of image feature matching, but they also have some limitations. Traditional algorithms such as the scale-invariant feature transform algorithm SIFT³⁶, the algorithm based on the fast orientation and rotation characteristics ORB³⁷, and the fast feature detection and description algorithm based on integral images SURF³⁸ are all detector-based methods. Newer methods like SuperPoint¹¹ leverage neural networks trained in an end-to-end fashion to perform image keypoint detection and feature description, while SuperGlue¹² utilizes graph neural networks and attention mechanisms to optimize the architecture for feature matching between image pairs.

Since detector-based methods rely heavily on the accuracy of keypoint detectors, this reliance introduces potential limitations. If detectors fail to accurately identify key points, significant features may be missed or incorrectly matched, especially in images with low texture, varying lighting conditions, or perspective changes. These methods tend to prioritize prominent local features, which may cause smaller but important details to be overlooked, leading to incomplete matches and reduced global consistency. As image data becomes increasingly complex, these limitations restrict further advancements in matching accuracy.

Detector-Free method

The detector-free approach bypasses keypoint detection, instead conducting full matching directly on feature maps of corresponding sizes. Methods like LOFTR¹⁹ employ self-attention and cross-attention mechanisms, while ASpanFormer²⁴ incorporates a hierarchical attention framework across various scales to strengthen global context comprehension and enhance the accuracy of feature matching. Other approaches, such as DISK³⁹, leverage probabilistic models combined with policy gradient training, and CAPS⁴⁰ employs a differentiable matching layer along with weak epipolar constraints to strengthen descriptor robustness. QuadTree⁴¹ reduces computational complexity with a token pyramid structure, and ASTR⁴² leverages a transformer-based decoder to avoid disturbances from irrelevant regions during feature aggregation. In parallel, dense pixel-level matching methods have gained attention for their ability to provide higher accuracy and a greater number of matches. For instance, methods like DKM²⁶, Roma²⁷, and PDCnet²⁸ adopt a two-stage process, starting with global matching followed by localized matching within smaller patches. Despite the increased computational cost, this approach significantly improves matching granularity, marking a major breakthrough in image matching.

Semantic Feature

Semantic feature extraction focuses on high-level, abstract information that conveys the meaning of an image or scene, making it essential for tasks such as scene understanding and object recognition. Unlike low-level features that detect textures or

edges, semantic features provide contextual insights, revealing the relationships between objects and characters in a scene. For instance, DeepLab⁴³ utilizes deep CNNs²⁹ to extract high-level semantic features and employs atrous convolution to enlarge the receptive field, capturing richer contextual information. Similarly, SegNet⁴⁴ introduces an encoder-decoder architecture, with the encoder capturing semantic features and the decoder progressively restoring spatial resolution. In contrast, PSPNet⁴⁵ leverages a multi-scale pyramid pooling mechanism to extract semantic features at different scales, thereby enhancing contextual understanding. In image matching tasks, we believe that semantic features, when combined with detailed low-level information, can enable more accurate comparisons by integrating both fine-grained details and high-level contextual understanding. This fusion is crucial for achieving precise and robust image matching.

However, most detector-free image matching methods overlook semantic information, focusing primarily on geometric and visual similarities. This reliance can limit their effectiveness in complex scenes with varying illumination, perspectives, or low texture. To overcome these limitations, we developed a semantic feature extraction module that dynamically captures key semantic content at different scales, alongside a semantic information enhancement module that further emphasizes these regions. By integrating these modules, our approach more effectively leverages semantic information, resulting in improved matching accuracy and performance, especially in challenging scenarios.

Methods

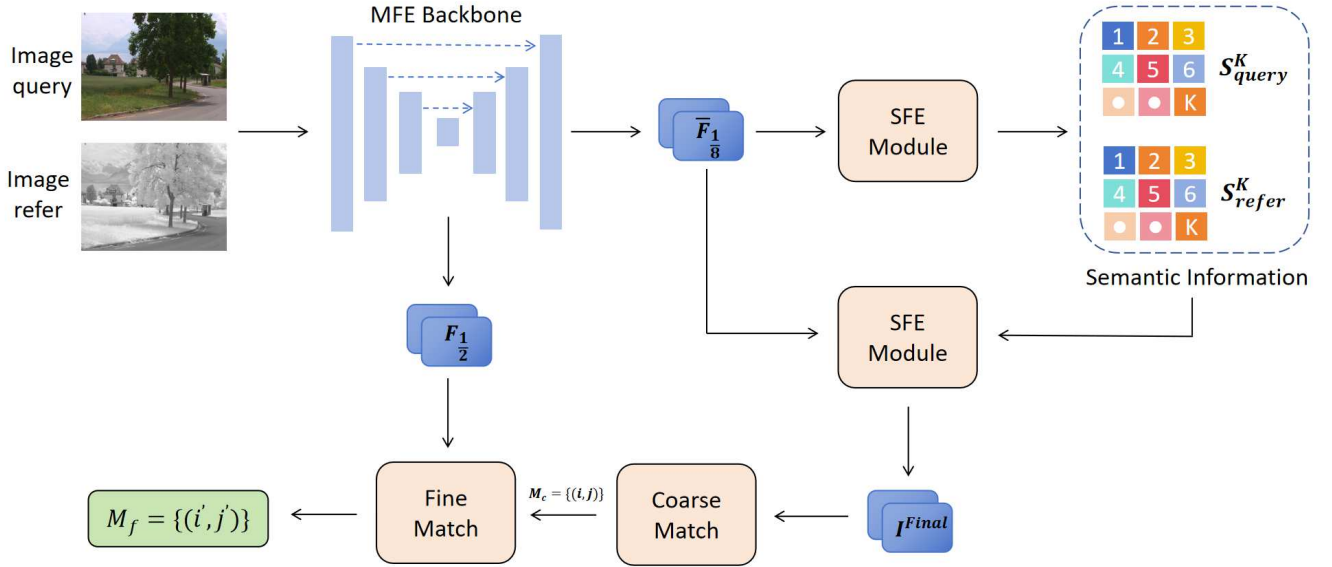


Figure 3. Overall structure diagram of SemFE

For the query image I_{query} and reference image I_{refer} , we first use a multi-scale feature extraction backbone that combines CNN⁴⁶ and Transformer⁴⁷ for feature extraction (Section MFE). Here, the feature map at 1/2 of the original resolution is used for fine matching, while the feature map at 1/8 of the original resolution is used for coarse matching. Next, the 1/8 scale feature map is sent to the semantic feature extraction and enhancement module (Section SFE; Section SIE). Coarse matching is then performed on the enhanced 1/8 scale feature map, following the same approach as FeMIP³⁵, and the results are further refined using the 1/2 scale feature map. Finally, both coarse and fine matching processes are supervised by the designed supervision method (Section Supervision). The structure of SemFE is illustrated in Fig 3.

Multi-scale Feature Extraction(MFE) backbone

In the feature extraction phase, the image is processed through a pyramid-structured backbone, similar to Unet⁴⁸, which integrates CNN⁴⁶ and Transformer⁴⁷. The backbone first reduces the image size while increasing the number of channels, followed by attention operations. It then performs upsampling and adds residual connections to extract feature maps at various scales. This process ultimately outputs a 1/8 scale feature map for coarse matching and a 1/2 scale feature map for fine matching. The detailed structure of the backbone is shown in (a) of Fig 4.

The formula for this part is as follows:

$$I_{1/16}, I_{1/8}, I_{1/4}, I_{1/2} = F_{CNN}(I) \quad (1)$$

$$M_{\frac{1}{16}} = SATT(I_{\frac{1}{16}} + PE) \quad (2)$$

Where I is the input image, $F_{CNN}(\cdot)$ represents the convolutional neural network layer, which consists of 4 layers. Each layer aims to reduce the image size and increase the number of channels, resulting in feature maps at different scales: $I_{\frac{1}{16}}, I_{\frac{1}{8}}, I_{\frac{1}{4}}, I_{\frac{1}{2}}$. $SATT(\cdot)$ denotes the self-attention operation, while PE stands for positional encoding. $M_{\frac{1}{16}}$ refers to a 1/16 scale feature map obtained through the self-attention mechanism. In addition, in the upsampling part, the model obtains three feature maps at varying scales. The formula flow of this part is as follows:

$$M_{\frac{1}{8}} = Up_{\frac{1}{8}}(M_{\frac{1}{16}}) + I_{\frac{1}{8}} \quad (3)$$

$$M_{\frac{1}{4}} = Up_{\frac{1}{4}}(M_{\frac{1}{8}}) + I_{\frac{1}{4}} \quad (4)$$

$$F_{\frac{1}{2}} = Up_{\frac{1}{2}}(M_{\frac{1}{4}}) + I_{\frac{1}{2}} \quad (5)$$

Where $Up(\cdot)$ represents the upsampling module, $M_{\frac{1}{8}}, M_{\frac{1}{4}}$ and $F_{\frac{1}{2}}$ represent the 1/8, 1/4 and 1/2 scale feature maps obtained by upsampling respectively. $M_{\frac{1}{8}}$ and $M_{\frac{1}{4}}$ are used for subsequent fusion, and $F_{\frac{1}{2}}$ is used for refined matching.

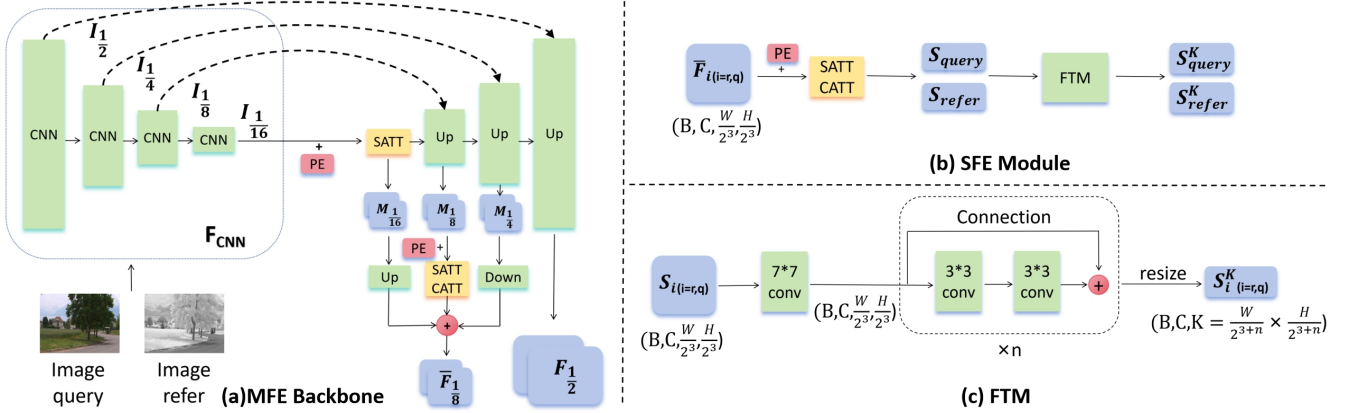


Figure 4. Detailed Structures of (a) MFE Backbone, (b) SFE Module and (c) FTM within the SemFE Framework

Next, self-attention and cross-attention will be applied to the 1/8 scale feature map $M_{\frac{1}{8}}$. Attention mechanisms are powerful tools for processing sequence or image data, which can establish effective connections between different positions or scales. The formulas are as follows:

$$F_{\frac{1}{8}} = SATT \& CATT(M_{\frac{1}{8}} + PE) \quad (6)$$

Where $CATT(\cdot)$ represents cross-attention operations, $F_{\frac{1}{8}}$ represents the feature map pair after attention mechanisms. Then, by formula (7), the feature map information of 1/4 and 1/16 scales will be fused into the feature map $F_{\frac{1}{8}}$.

$$\bar{F}_{\frac{1}{8}} = F_{\frac{1}{8}} + Up(M_{\frac{1}{16}}) + Down(M_{\frac{1}{4}}) \quad (7)$$

Where $Down(\cdot)$ represents downsampling and $\bar{F}_{\frac{1}{8}}$ represents the feature map after fusion.

Finally, the backbone outputs $\bar{F}_{\frac{1}{8}}$ and $F_{\frac{1}{2}}$. $\bar{F}_{\frac{1}{8}}$, the multi-scale fused feature map, is used for both semantic information extraction and enhancement, and is simplified as $\bar{F}$ in subsequent operations. $F_{\frac{1}{2}}$ represents the 1/2 scale feature map, which is employed for fine matching.

This fusion process organically combines information at different scales, fully utilizes the feature expressions at each scale, helps the model understand and process image data more comprehensively, and achieves more accurate and robust results in matching tasks. By fusing feature information at different scales, our aim is to enhance the model's capacity to discern image subtleties, thereby achieving better performance in image matching tasks.

Semantic Feature Extraction(SFE) Module

The SFE module is responsible for capturing semantic information for subsequent enhancement. In this module, the feature map is processed through a carefully designed local semantic extraction system, composed of attention mechanisms and a feature transformation module (FTM), as illustrated in (b) of Fig 4. This module extracts K representative local semantic features, effectively dividing the feature map into K distinct semantic regions, each representing a generalized feature representation.

First, two different attention mechanisms are implemented on the feature map, along with the addition of positional encoding. The self-attention mechanism allows the model to focus more on critical regions and information within the feature map during processing, thereby achieving effective interaction and fusion of information within the feature map. The cross-attention mechanism further promotes the exchange and fusion of information between different feature maps, helping the model to establish closer and more effective associations between feature maps. These two steps lay a solid foundation for the subsequent feature fusion stage, allowing the model to more effectively comprehend and leverage the information contained in the feature map. In addition, the inclusion of position encoding offers the model crucial information about the spatial locations of elements within the feature map, allowing the model to consider the key factor of spatial position when understanding the feature map. The formula for this step is as follows:

$$S_{i(i=r,q)} = SATT \& CATT(\bar{F}_{i(i=r,q)} + PE) \quad (8)$$

Where S_i represents the features of the processed image pair, and PE denotes positional encoding. $SATT(\cdot)$ and $CATT(\cdot)$ represent attention operations. Following this, the local semantic feature extraction module extracts K semantic features from the feature map. The number K is dynamically established according to the size and complexity of the feature map, as we assume that larger and more information-rich maps require more semantic features for a comprehensive representation. Conversely, smaller or simpler feature maps demand fewer semantic features. This dynamic adjustment ensures that the most representative local semantic features are extracted. The formula for K is as follows:

$$K = \left\lceil \frac{W}{2^{3+n}} \right\rceil \times \left\lceil \frac{H}{2^{3+n}} \right\rceil, n = 3 \quad (9)$$

In the process of local semantic feature extraction, K is a key parameter determined by n. After extensive experiments and fine-tuning, n is set as a hyperparameter with a value of 3. In this way, the parameter K will be dynamically adjusted based on the size and complexity of the input image. Specifically, the width (W) and height (H) of the input image directly affect the number and distribution of local semantic features.

The feature transformation module, composed of a series of one-dimensional convolutions (as shown in (c) of Fig 4), further refines the extracted local semantic features. This module outputs two local semantic feature maps with dimensions (B,C,K), where K represents the amount of semantic information. These maps encapsulate highly generalized and abstract representations of the image content, offering both discriminative and representative features. These features not only distinguish different image elements but also effectively capture the most critical semantic information within the image. The process is summarized in the following formula:

$$S_{query}^K = FTM(S_{query}) \quad (10)$$

$$S_{refer}^K = FTM(S_{refer}) \quad (11)$$

Where S_{query} and S_{refer} represent the query image and reference image in the feature map, respectively. $FTM(\cdot)$ is feature transformation module. S_{query}^K and S_{refer}^K represent the local semantic feature maps of the query and reference images, respectively.

Semantic Information Enhancement(SIE) Module

The SIE module aims to enhance the feature map by leveraging semantic information. In this module, the 1/8 scale feature map is first compared with the feature map containing K local semantic features to compute similarity (see Fig 5(a) for the flowchart). Each token in the 1/8 scale feature map is then assigned to one of the K local semantic groups based on this similarity. Subsequently, self-attention operations are performed within each of the K groups. This requires K separate self-attention processes, each promoting interaction and fusion between features that share the same semantic group. This step enriches the feature representation, enhancing both the distinctiveness and expressiveness of the feature map. The process of feature similarity calculation, grouping, and attention operation is formulated as follows:

$$G_{query}^K = Group(\bar{F}_{query}, Similarity(\bar{F}_{query}, S_{query}^K)) \quad (12)$$

$$G_{refer}^K = \text{Group}(\bar{F}_{refer}, \text{Similarity}(\bar{F}_{refer}, S_{refer}^K)) \quad (13)$$

$$\bar{G}_{query} = \{\bar{G}_{query}^i\}_{i=1}^K = \{\text{SATT}(G_{query}^i + PE)\}_{i=1}^K \quad (14)$$

$$\bar{G}_{refer} = \{\bar{G}_{refer}^i\}_{i=1}^K = \{\text{SATT}(G_{refer}^i + PE)\}_{i=1}^K \quad (15)$$

Where $\text{Similarity}(\cdot)$ is the calculation of similarity, which is calculated using cosine similarity. $\text{Group}(\cdot)$ is to assign the feature $\bar{F}_{query}$ to different groups according to similarity. The purpose is to classify each token in the feature map into which group. G_{query}^K and G_{refer}^K represent the query feature map and refer feature map that have been grouped, respectively. $\bar{G}_{query}$ and $\bar{G}_{refer}$ are the feature maps of all groups after self-attention. $\bar{G}_{query}^i$ and $\bar{G}_{refer}^i$ represent the features of the i -th group of features after self-attention. G_{query}^i and G_{refer}^i represent the i -th group of features in the K groups of features.

Through these operations, two 1/8 scale feature maps, $\bar{G}_{query}$ and $\bar{G}_{refer}$, are generated. These maps encapsulate multi-level and multi-stage processing and interactions, resulting in richer feature representations and a deeper understanding of image content. This step aims to enhance the model's adaptability and accuracy in image matching tasks, providing a stronger foundation for subsequent matching processes.

In the next phase, the 1/8 scale feature maps $\bar{G}_{query}$ and $\bar{G}_{refer}$ undergo cross-attention operations with the local semantic feature maps S_{query}^K and S_{refer}^K . Here, the semantic information S^K acts as the query (Q), while the feature maps $\bar{G}$ serve as the key and value (K, V). The effect of this step is that each token in S^K is asking which positions or features in $\bar{G}$ are most important or relevant to it. This mechanism allows the model to focus on specific areas in $\bar{G}$ based on semantic information, and use semantic information to improve the feature interaction process, thus strengthening the model's ability to interpret the semantic content of the image and improving the effectiveness and accuracy of image matching. Fig 5(a) illustrates the semantic information enhancement module, and the corresponding formulas for this operation are as follows:

$$\bar{S}_{query}^K = \text{CATT}(Q = S_{query}^K, K = \bar{G}_{query}, V = \bar{G}_{query}) \quad (16)$$

$$\bar{S}_{refer}^K = \text{CATT}(Q = S_{refer}^K, K = \bar{G}_{refer}, V = \bar{G}_{refer}) \quad (17)$$

Where the query (Q) in the cross attention comes from the local semantic feature maps S_{query}^K and S_{refer}^K , while the key (K) and value (V) come from the feature maps $\bar{G}_{query}$ and $\bar{G}_{refer}$. $\bar{S}_{query}^K$ and $\bar{S}_{refer}^K$ represent the feature maps after completing cross attention.

Next, the feature map $\bar{G}_{query}$ and $\bar{G}_{refer}$ serve as the query (Q), while the local semantic feature maps $\bar{S}_{query}^K$ and $\bar{S}_{refer}^K$ provide the key (K) and the value (V). In this step, the low-level features in $\bar{G}$ query the semantic information related to them in $\bar{S}^K$ through the cross-attention mechanism. By using the semantic features in $\bar{S}^K$ as the key and value, the network can match the semantic information for each local feature of $\bar{G}$, thereby improving the semantic consistency and expressiveness of the features. This step is mainly to allow ordinary spatial and texture features to benefit from the guidance of semantic features, further improving the distinguishing ability and overall expressiveness of the features. The specific formula is as follows:

$$\dot{G}_{query} = \text{CATT}(Q = \bar{G}_{query}, K = \bar{S}_{query}^K, V = \bar{S}_{query}^K) \quad (18)$$

$$\dot{G}_{refer} = \text{CATT}(Q = \bar{G}_{refer}, K = \bar{S}_{refer}^K, V = \bar{S}_{refer}^K) \quad (19)$$

Where $\dot{G}_{query}$ and $\dot{G}_{refer}$ represent feature maps after this cross attention. Then the cross attention will be applied to the feature maps $\dot{G}_{query}$ and $\dot{G}_{refer}$, so that they further interact with each other and fuse the features of the two images to obtain the feature map used for the final matching. The specific formula is as follows:

$$I^{Final} = \text{SATT\&CATT}(\dot{G}_{query}, \dot{G}_{refer}) \quad (20)$$

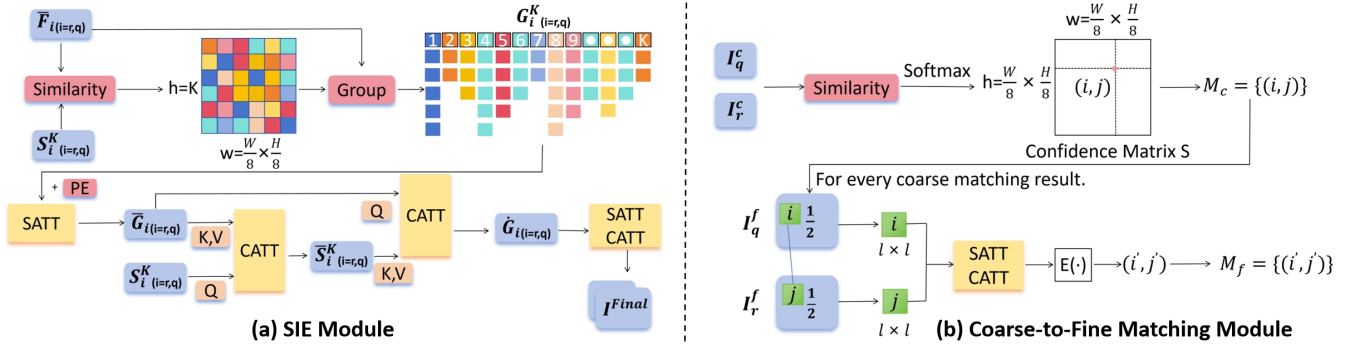


Figure 5. Detailed Structures of (a) SIE Module and (b) Coarse-to-Fine Matching Module within the SemFE Framework

$SATT(\cdot)$ and $CATT(\cdot)$ represent attention operations, and I^{Final} represents the feature map pairs to be matched in the end.

Finally, after these attention mechanism operations, the semantic information enhancement module outputs two final feature map pairs, I^{Final} . These feature maps, having undergone multiple high-level interactions and fusions, possess richer and more comprehensive representational capabilities. They serve as the key feature representations for subsequent image matching tasks, significantly improving the model's accuracy. This multi-level feature fusion, enabled by the attention mechanism, is crucial to the success of image matching.

Coarse-to-Fine Matching Module

The Coarse-to-Fine Matching Module is designed to achieve robust and accurate matching between features through a two-stage process: coarse matching followed by fine matching. The flowchart of coarse matching and fine matching is shown in (b) of Fig 5.

Coarse Matching

In the coarse matching stage, we define I_q^c and I_r^c as the query feature map and reference feature map of the $1/8$ scale feature map pair I^{Final} . A series of corresponding relationships are constructed through I_q^c and I_r^c to evaluate the matching degree between different tokens. The matching probabilities between elements are quantified using a bidirectional softmax operation. Specifically, the matching probability $P(i \leftrightarrow j)$ is calculated from feature I_q^c to feature I_r^c and vice versa. These probabilities are normalized using the softmax function to guarantee that the total probability in each direction sums to 1. This normalization provides a reliable basis for applying reward and penalty strategies during training. The formula is as follows:

$$P(i \leftrightarrow j) = \frac{\text{softmax}(S(i, \cdot))j \cdot \text{softmax}(S(\cdot, j))i}{\sum_{i^m, j^n} \text{softmax}(S(i^m, \cdot))j^n \cdot \text{softmax}(S(\cdot, j^n))i^m} \quad (21)$$

In this formula, m represents all row indices of the confidence matrix S . n represents all column indices of the confidence matrix S . The confidence matrix S is used to establish correspondences between feature points and filter out unreliable ones. Finally, a series of matching points M_c with coarse matching will be obtained.

Fine Matching

Coarse matching provides fast but approximate matching, while the fine matching stage improves this process to achieve higher pixel-level accuracy. In the fine matching stage, we define I_q^f and I_r^f as the query feature map and reference feature map of the $1/2$ scale feature map pair $F_{1/2}$. The fine matching stage first maps the coordinates of candidate matching points from the coarse feature map to higher resolution fine-grained feature maps I_q^f and I_r^f . Around these mapped points, local feature windows are extracted to capture both coarse and fine details. These windows are then fused to create combined feature maps I_q^{fc} and I_r^{fc} , which are input into the interaction module containing two attention mechanisms. This module strengthens the feature representation and deepens the relationship between features and finally the precise matching result M_f is obtained by calculating the expected value.

Supervision

Coarse Matching Loss

During the coarse matching process, the supervised method adopts the same policy gradient method in FeMIP³⁵, combining reward and penalty strategies to optimize model performance. A threshold u is employed, with dynamically adjusted rewards and penalties guiding the coarse matching training process. Specifically, a positive reward α is granted when the highest

matching probability of a positive sample exceeds the threshold, while a penalty is applied if this criterion is not met, or if a negative sample incorrectly surpasses the threshold. To balance the strength of rewards and penalties, the negative reward β is dynamically adjusted during training. The adjustment method for β is as follows:

$$\beta = \begin{cases} 0, & \text{if epoch} < n \\ -0.01 \cdot (\text{epoch} - n), & \text{if } n \leq \text{epoch} < n + 25 \\ -0.25, & \text{if epoch} \geq n + 25 \end{cases} \quad (22)$$

The coarse matching loss L_1 is derived from the policy gradient approach, which optimizes the matching process by minimizing the mismatch between features I_q^c and I_r^c . The loss function incorporates the matching degree distribution $P(i \leftrightarrow j | I_q^c, I_r^c)$ and the logarithmic gradient $\nabla_{\theta \phi_{ij}}$ of the action sequence. This loss directly reflects the quality of the match and helps the model learn to extract and align similar features more effectively.

The expression of policy gradient is as follows:

$$\nabla_{\theta} E_{M_{q \leftrightarrow r}} R_{(M_{q \leftrightarrow r})} = E_{I_q^c, I_r^c} \sum_{i,j} [P(i \leftrightarrow j | I_q^c, I_r^c) \cdot r(i \leftrightarrow j) \cdot \nabla_{\theta \phi_{ij}}] \quad (23)$$

$$\phi_{ij} = \log P(i \leftrightarrow j | I_q^c, I_r^c) + \log P(I_q^c, i | I_q^c, \theta_x) + \log P(I_r^c, j | I_r^c, \theta_x) \quad (24)$$

Finally, the loss function of the coarse matching level is as follows, and t is the given confidence threshold:

$$L_1 = \nabla_{\theta} E_{M_{q \leftrightarrow r}} R_{(M_{q \leftrightarrow r})}, p(i \leftrightarrow j) > t \quad (25)$$

Fine Matching Loss

The fine matching loss L_2 is defined for the refinement level regression and re-refinement modules. This step involves calculating the slight differences $(\nabla x, \nabla y)$ between the predicted and actual coordinates to measure the precision of the matching, it aims to minimize the error between the predicted coordinates $(i + \nabla x, j + \nabla y)$ and the actual coordinates (i_a, j_a) . The loss is averaged over all m feature points involved in the matching process. By minimizing this error, the model can progressively refine the precision of its predictions. The L_2 loss function is expressed as follows:

$$L_2 = \frac{1}{m} \sum_{i=1}^m (i + \nabla x, j + \nabla y) - (i_a, j_a)^2 \quad (26)$$

Total Loss Function

The total loss function combines the losses from both the coarse and fine matching stages, weighted by factors δ_1 and δ_2 , respectively:

$$L_{total} = \delta_1 L_1 + \delta_2 L_2 \quad (27)$$

This combined loss function guides the overall training process, ensuring that both coarse and fine matching stages contribute to the final model performance.

Experiment

Dataset

To assess the broad applicability and effectiveness of our model, we carefully selected five multimodal datasets with different characteristics. These datasets not only cover a variety of scenes from natural landscapes to urban environments, but also contain a rich variety of image modal types, such as RGB, depth, optical, near-infrared (NIR), synthetic aperture radar (SAR), and short-wave infrared (SWIR) images.

(1) **SEN12MS dataset:** SEN12MS provides a rich combination of image modalities, including RGB, SAR, NIR, and SWIR images. We selected 16,000 image pairs for the model's training phase to ensure it fully captures the associations and distinctions between the modalities. Additionally, to assess the model's generalization ability, we set aside 600 image pairs for testing.

(2) **Optical-SAR dataset:** The Optical-SAR dataset shows a variety of complex scenes such as plains, harbors, cities, islands, and rivers. Each pair of images contains both optical and SAR modalities, and the size is unified to 512×512. We allocated 16,000 image pairs for training, while another 500 pairs were set aside for testing to assess the model's effectiveness in real-world scenarios.

Table 1. Homography estimation results on the SEN12MS dataset.

Different Datasets	Method	Homography est. AUC		
		@3px	@5px	@10px
NIR-RGB	MatchNet (2015)	29.49	49.61	71.85
	Tfeat (2016)	45.22	60.28	75.31
	HardNet (2017)	64.44	76.37	88.18
	MatchosNet (2021)	63.64	77.32	87.92
	LoFTR (2021)	63.69	76.71	87.69
	LNIFT (2022)	53.04	66.22	78.79
	FeMIP (2023)	61.30	77.93	88.21
	FmCFA (2023)	65.96	79.33	89.66
	SemFE (Ours)	66.39	79.83	89.91
SAR-SWIR	MatchNet (2015)	0.16	0.73	1.13
	Tfeat (2016)	0.34	0.92	2.17
	HardNet (2017)	7.33	20.85	42.54
	MatchosNet (2021)	8.51	21.90	42.75
	LoFTR (2021)	32.37	49.63	67.41
	LNIFT (2022)	1.07	2.38	3.44
	FeMIP (2023)	28.26	43.09	70.69
	FmCFA (2023)	62.43	76.79	88.37
	SemFE (Ours)	66.20	79.72	89.86
SAR-NIR	MatchNet (2015)	0.11	0.36	1.24
	Tfeat (2016)	0.57	1.66	4.74
	HardNet (2017)	10.08	24.77	47.60
	MatchosNet (2021)	10.47	25.62	48.52
	LoFTR (2021)	39.72	53.17	75.67
	LNIFT (2022)	1.46	3.12	5.09
	FeMIP (2023)	23.19	53.26	76.15
	FmCFA (2023)	42.45	65.40	82.70
	SemFE (Ours)	51.37	70.82	85.41
NIR-SWIR	MatchNet (2015)	10.87	21.51	37.28
	Tfeat (2016)	29.03	42.60	56.06
	HardNet (2017)	51.54	68.85	82.94
	MatchosNet (2021)	50.48	68.19	52.56
	LoFTR (2021)	61.98	75.50	86.97
	LNIFT (2022)	49.10	65.04	78.99
	FeMIP (2023)	66.45	79.77	89.89
	FmCFA (2023)	69.70	81.22	90.56
	SemFE (Ours)	69.98	81.63	90.81

(3) **WHU-OPT-SAR dataset:** This dataset focuses on the diverse geographical features of Hubei Province, China. The WHU-OPT-SAR dataset covers six categories: forest, road, water, village, farmland, and city. Since the original images were large, we carefully cropped them and finally generated 4,400 image pairs. Among them, 3,800 pairs of images were allocated at random to the training set to promote model training; the other 300 pairs of images were used for verification and testing to ensure that the model can be fully evaluated at different stages.

(4) **RGB-NIR scene dataset:** This dataset focuses on the integration of RGB and NIR images, aiming to explore the complementarity of these two modalities in scene understanding. It covers nine typical scenes such as villages, streets, fields, and forests, and contains a total of 477 pairs of images. We meticulously curated 400 pairs of images as training data to train the model to recognize RGB and NIR features in different scenes; at the same time, 48 pairs of images were retained as test sets to verify the recognition accuracy of the model in unknown scenes.

(5) **NYU-Depth V2 dataset:** For indoor environments, NYU-Depth V2 provides RGB and depth image pairs captured from video sequences with dense annotations, totaling 1449 pairs. We divided 1,049 pairs of images for model training to strengthen the model's grasp of indoor spatial structures and their depth information; the remaining 400 pairs of images served as a test set to thoroughly assess the model's effectiveness in indoor scenes.

Table 2. Homography Estimation on Multimodal Datasets.

Different Datasets	Method	Homography est. AUC		
		@3px	@5px	@10px
Optical-SAR	MatchNet (2015)	3.78	14.24	38.04
	Tfeat (2016)	12.74	31.06	56.39
	HardNet (2017)	24.19	45.27	67.89
	MatchosNet (2021)	20.87	42.36	65.38
	LoFTR (2021)	29.54	45.66	64.86
	FeMIP (2023)	28.89	49.4	70.07
	FmCFA (2023)	53.81	67.62	81.91
	SemFE(Ours)	67.65	79.95	89.96
RGB-NIR Scene	MatchNet (2015)	0.61	3.51	15.35
	Tfeat (2016)	0.42	2.39	12.33
	HardNet (2017)	18.35	39.02	63.37
	MatchosNet (2021)	12.54	28.8	52.09
	LoFTR (2021)	41.66	60.66	78.26
	FeMIP (2023)	37.56	61.74	78.65
	FmCFA (2023)	43.05	62.39	80.56
	SemFE (Ours)	55.47	68.96	83.35
WHU-OPT-SAR	MatchNet (2015)	3.83	14.39	32.46
	Tfeat (2016)	2.52	13.47	30.06
	HardNet (2017)	16.51	39.1	66.25
	MatchosNet (2021)	14.72	33.2	58.81
	LoFTR (2021)	17.4	41.98	67.57
	FeMIP (2023)	21.72	44.7	69.75
	FmCFA (2023)	43.51	57.18	75.82
	SemFE (Ours)	58.71	74.02	86.27
NYU-Depth V2	MatchNet (2015)	0.61	3.51	15.35
	Tfeat (2016)	0.42	2.39	12.33
	HardNet (2017)	18.35	39.02	63.37
	MatchosNet (2021)	12.54	28.8	52.09
	LoFTR (2021)	61.89	76.04	88.03
	FeMIP (2023)	69.08	80.36	89.72
	FmCFA (2023)	74.64	84.40	92.20
	SemFE (Ours)	79.77	87.45	93.70

Implementation Details

During the training of the model, we employed the AdamW⁴⁹ optimizer along with the CosineAnnealingLR⁵⁰ learning rate scheduling strategy to enhance model performance. Specifically, the learning rate was set to $1 * 10^{-4}$. At the same time, in order to balance memory usage and training efficiency, we configured the batch size as 4 and established the threshold value t at 0.2. In addition, we predefined two parameters, δ_1 and δ_2 , both set to 1. These parameters may be used to adjust the learning rate, increase the regularization strength, or trigger specific training logic. Nvidia RTX 3090 GPU is used for the entire training process.

Evaluation Indicators

To demonstrate the effectiveness of our design method in multimodal image feature matching, we carefully planned a comparative experiment with the most advanced methods in the field, covering outstanding algorithms such as FeMIP³⁵, LoFTR¹⁹, HardNet⁵¹, TFeat⁵², MatchosNet⁵³, MatchNet⁵⁴, and FmCFA. All the methods involved in the comparison were trained with the same training data set under a unified environment configuration, and performance tests were carried out on a consistent test set to ensure fairness and accuracy.

In the evaluation phase, we used dual evaluation indicators to comprehensively measure the performance of each method: Homography estimation and robustness verification: The homography matrix is calculated using the powerful image processing capabilities of OpenCV, and the RANSAC algorithm is then introduced to improve the reliability of the estimation process. Specifically, we ascertain valid matches by computing the angular discrepancy between the predicted and actual values, with particular attention paid to the mean discrepancy of the four corners upon image reprojection. On this basis, we present the

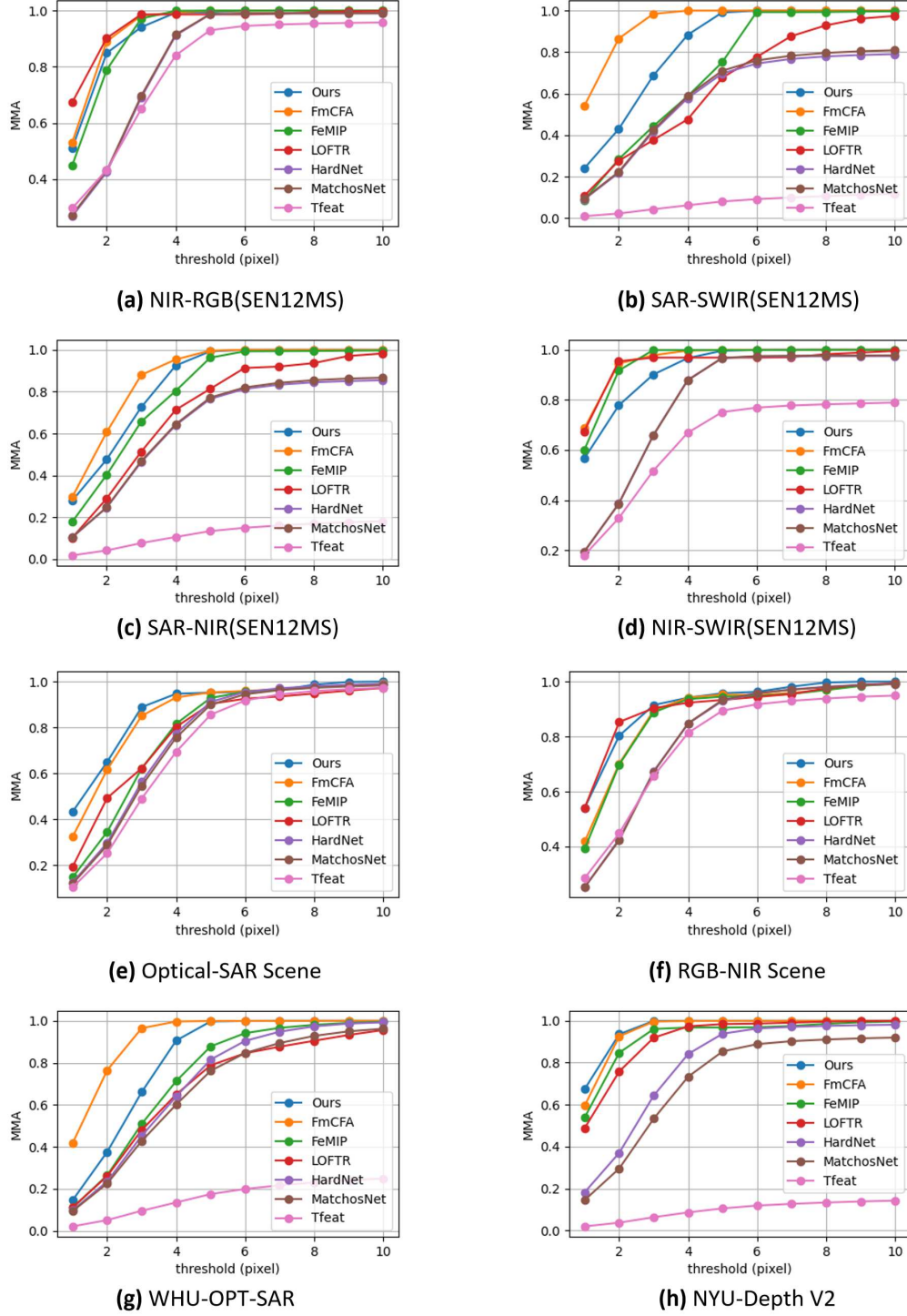


Figure 6. Average matching accuracy experiments. (a) NIR RGB, (b) SAR-SWIR, (c) SAR-NIR, (d) NIR-SWIR, (e) Optical-SAR Scene, (f) RGB-NIR Scene, (g) WHU-OPT-SAR, (h) NYU-Depth V2.

aggregate area under the curve (AUC) values at three crucial error thresholds (@3 pixels, @5 pixels, @10 pixels), which intuitively reflects the comprehensive performance of the method under different accuracy requirements.

Mean Matching Accuracy (MMA) evaluation: To gain insight into the quality of feature matching, we introduced MMA as another core evaluation metric. MMA focuses on evaluating the closeness of feature matching between image pairs, and only considers matching pairs that are nearest neighbors to each other. During the evaluation process, we set a series of pixel

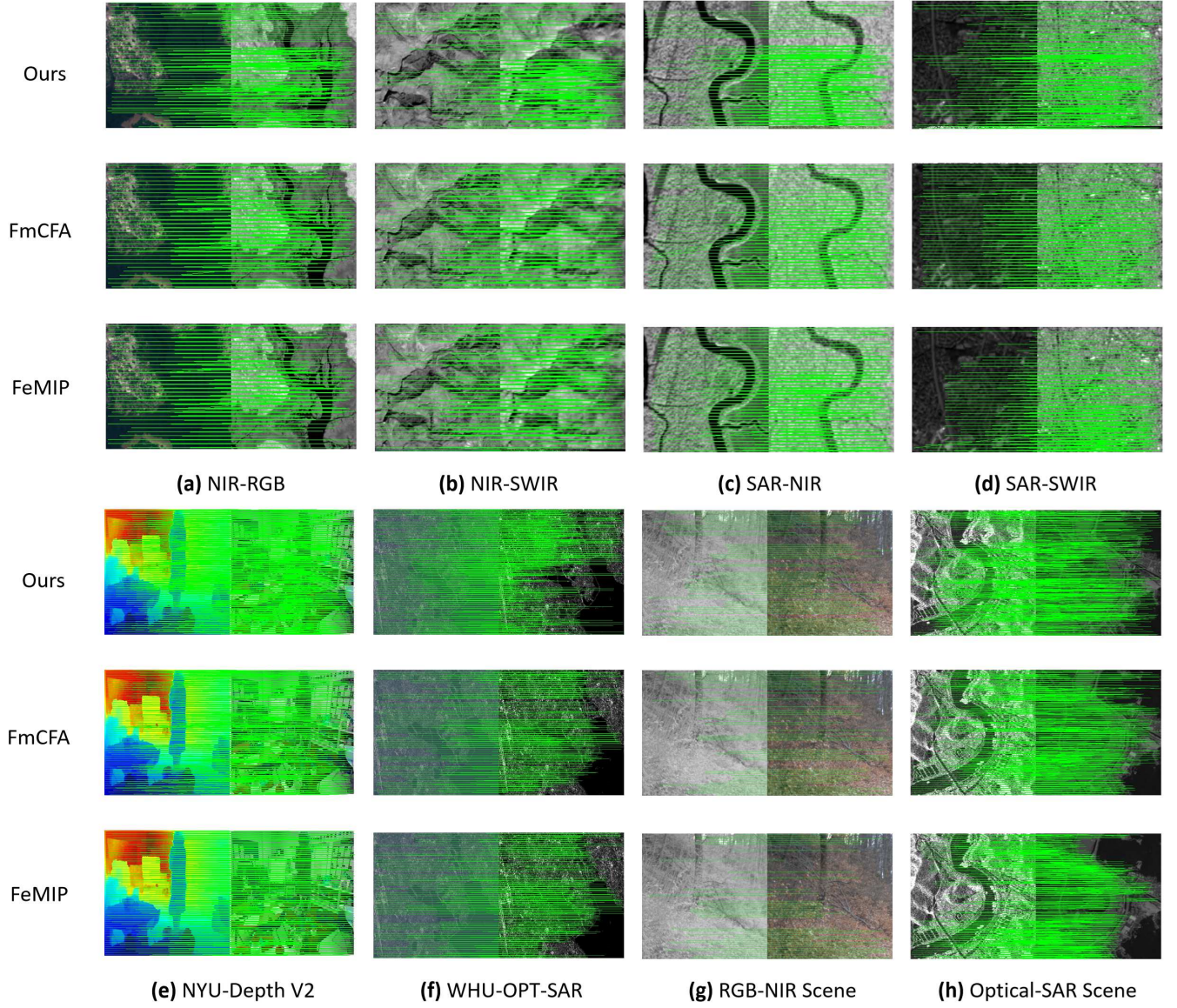


Figure 7. Feature matching visualization under different datasets. (a) NIR-RGB, (b) NIR-SWIR, (c) SAR-NIR, (d) SAR-SWIR, (e) NYU-Depth V2, (f) WHU-OPT-SAR, (g) RGB-NIR Scene, (h) Optical-SAR Scene.

matching thresholds and calculate the percentage of correct matches based on these thresholds. A correct match is defined as one whose homography reprojection error is lower than the set matching threshold. Finally, MMA comprehensively evaluates the matching accuracy of the method on all image pairs by calculating the weighted average of the average correct match percentages under each threshold, providing a quantitative basis for judging the efficacy of the method.

Homography Estimation Experiment

In the experimental framework of homography estimation, we focus on performance evaluation under multiple angular error thresholds (3, 5, and 10 pixels), and use AUC as a key indicator to measure the accuracy of each method. For the SEN12MS data set, we carefully selected four representative groups of modal image data pairs (including NIR-RGB, NIR-SWIR, SAR-NIR, and SAR-SWIR) for detailed testing. The results of our experiments demonstrate that the SemFE methods we proposed all exhibit outstanding performance, not only leading in accuracy, but also reflecting its powerful cross-modal matching capabilities. The results are shown in Table 1.

To further assess the generalization capability of SemFE, we also extended the experimental scope to multiple data sets such as Optical-SAR, RGB-NIR, WHU-OPT-SAR, and NYU-Depth V2. On all these datasets, SemFE consistently demonstrates significant advantages over other contrasting methods, whether under strict threshold conditions of 3, 5, or 10 pixels. These

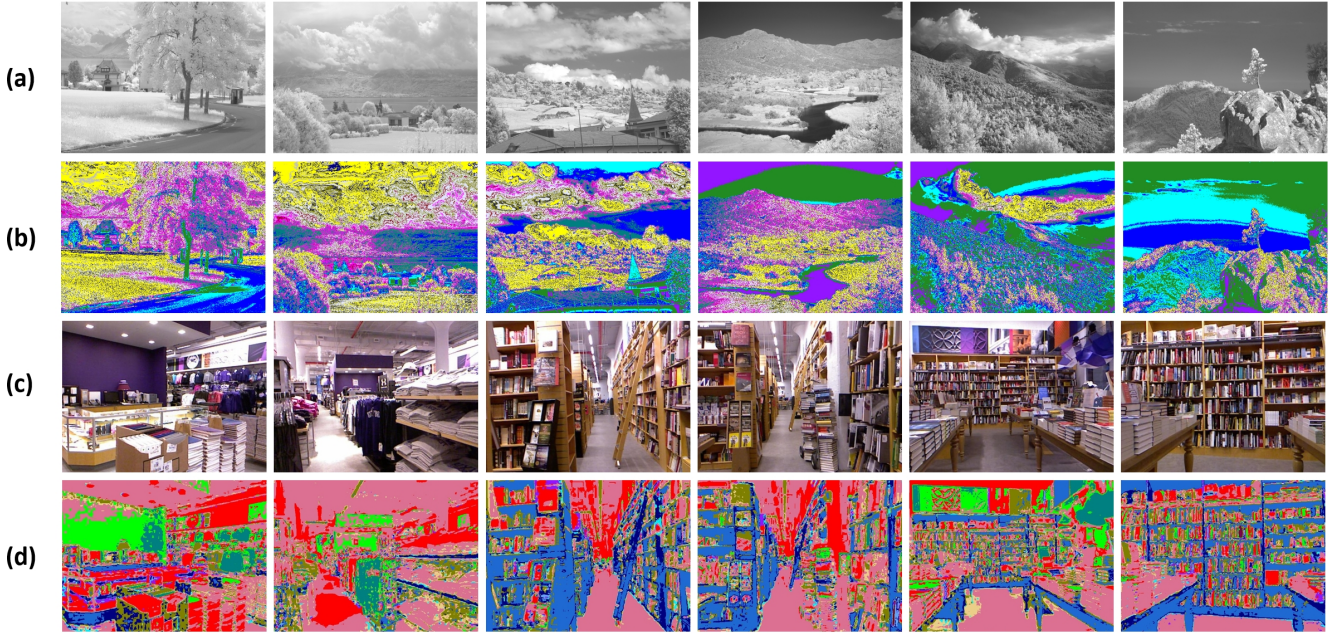


Figure 8. NIR Modality Data (a) and Corresponding Semantic Information Assignment (b) from the RGB-NIR Scene Dataset, and RGB Modality Data (c) with Corresponding Semantic Information Assignment (d) from the NYU-Depth V2 Dataset.

results demonstrate the efficiency and accuracy of SemFE in the homography estimation task. The results are shown in Table 2.

Matching Accuracy Experiment

In the average matching accuracy experiment, we specifically selected four challenging modality image pairs (NIR-RGB, NIR-SWIR, SAR-NIR, and SAR-SWIR) from the SEN12MS dataset as test cases. Additionally, we extended the experiment to several other modality datasets, including the RGB-NIR Scene dataset, NYU-Depth V2, Optical-SAR, and WHU-OPT-SAR. The experimental results are shown in Fig 6, where the x-axis depicts the variation in pixel threshold, and the y-axis reflects the proportion of correct matches. From the Fig 6, it is evident that methods such as SemFE, FmCFA, FeMIP³⁵, and LoFTR¹⁹ demonstrate strong performance, standing out among many competitors. Notably, SemFE achieves consistently high matching accuracy across most scenarios.

Multimodal Image Matching Visualization

In the comparative evaluation, we directly compared SemFE with the FeMIP³⁵ and FmCFA methods to verify its advantages in multimodal image matching. For the SEN12MS dataset, we selected four representative modality image pairs: NIR-RGB, SAR-SWIR, SAR-NIR, and NIR-SWIR. Additionally, we applied SemFE to four other datasets of different types: NYU Depth V2, Optical-SAR, RGB-NIR Scene, and WHU-OPT-SAR. The matching performance of each method is illustrated in Fig 7. The experimental results show that SemFE not only generates more matching connections but also achieves higher accuracy than FmCFA and FeMIP³⁵, demonstrating its superior capability in complex multimodal image matching tasks.

Semantic Information Visualization Experiment

In this experiment, we visualize the semantic information extracted by our SemFE method using two datasets: RGB-NIR Scene (Fig 8(a)) and NYU-Depth V2 (Fig 8(c)). The aim is to demonstrate the model’s effectiveness in enhancing and assigning semantic information across different modalities. Fig 8(b) shows the NIR modality results from the RGB-NIR Scene dataset, where our method accurately captures key semantic features that are crucial for the matching process. Fig 8(d) displays the RGB modality results from the NYU-Depth V2 dataset, where SemFE effectively extracts and enhances semantic information, adapting to varying scene depths and complex structures. These visualizations demonstrate that SemFE significantly improves semantic feature extraction, leading to more accurate feature matching and better overall performance across multimodal datasets.

Ablation Experiment

We performed a series of ablation experiments to assess the effectiveness of key modules in the homography estimation task on the NYU-Depth V2 dataset. In these experiments, we studied three key modules: Multi-Scale, Attention-Backbone, and

Table 3. Homography Estimation on Multimodal Datasets.

Dataset	Module			Homography est. AUC		
	Multi-Scale	Attention-Backbone	SIE-Module	@3px	@5px	@10px
NYU-Depth V2	×	×	×	64.69	78.48	89.24
	✓	×	×	66.76	79.66	89.85
	×	✓	×	69.96	81.64	90.82
	×	×	✓	76.96	85.80	92.90
	✓	✓	×	73.44	83.62	91.81
	✓	×	✓	76.73	86.02	93.01
	×	✓	✓	78.18	86.44	93.19
	✓	✓	✓	79.77	87.45	93.70

SIE-Module. Multi-Scale represents the multi-scale feature extraction method used in MFE, Attention-Backbone represents the attention mechanism in MFE. SIE-Module is the semantic information enhancement module, which is our main innovative module. We measured the performance by evaluating the accuracy of the model at different thresholds (@3px, @5px, @10px). The results of the experiments are presented in Table 3.

The experimental results indicate that the incorporation of the SIE-Module substantially enhances the accuracy of homography estimation, especially at lower thresholds. At the same time, the introduction of the Multi-Scale and Attention-Backbone contributes to enhancing the model's performance, but the improvement in accuracy is relatively small.

Conclusion

This paper introduces an innovative feature matching method for multimodal image datasets, referred to as SemFE. This method employs a multi-scale feature extraction backbone network combined with an attention mechanism. Additionally, we designed a semantic feature extraction module that can dynamically extract semantic information based on the image size. We also developed a novel semantic information enhancement module that enhances image features according to the semantic information. Numerous experiments have been conducted on various modal datasets, and SemFE has achieved excellent matching results, further proving its effectiveness and robustness.

Data Availability

The datasets examined in this study can be accessed from the following publicly available resources.

SEN12MS dataset: <https://opendatalab.com/OpenDataLab/SEN12MS>

Optical-SAR dataset: <https://www.kaggle.com/datasets/requiemonk/sentinel12-image-pairs-segregated-by-terrain/data>

WHU-OPT-SAR dataset: <https://github.com/AmberHen/WHU-OPT-SAR-dataset>

RGB-NIR scene dataset: <https://www.epfl.ch/labs/ivrl/research/downloads/rgb-nir-scene-dataset/>

NYU-Depth V2 dataset: <https://www.kaggle.com/datasets/soumikrakshit/nyu-depth-v2>

Code availability

The computer algorithm code has been made public and uploaded to github.<https://github.com/TengRui/SemFE>

References

1. Wang, Z. Automatic and robust hand gesture recognition by sdd features based model matching. *Appl. Intell.* **52**, 11288–11299 (2022).
2. Lai, B. *et al.* 2d3d-mvpnet: Learning cross-domain feature descriptors for 2d-3d matching based on multi-view projections of point clouds. *Appl. Intell.* **52**, 14178–14193 (2022).
3. Yang, L., Huang, Q., Li, X. & Yuan, Y. Dynamic-scale grid structure with weighted-scoring strategy for fast feature matching. *Appl. Intell.* **52**, 10576–10590, DOI: [10.1007/s10489-021-02990-3](https://doi.org/10.1007/s10489-021-02990-3) (2022).
4. Ma, J., Fan, A., Jiang, X. & Xiao, G. Feature matching via motion-consistency driven probabilistic graphical model. *Int. J. Comput. Vis.* **130**, 2249–2264, DOI: [10.1007/s11263-022-01644-2](https://doi.org/10.1007/s11263-022-01644-2) (2022).
5. Fan, X. *et al.* Vls-g-sanet: a feature matching algorithm for remote sensing image registration. *Knowledge-Based Syst.* **255**, 109609, DOI: [10.1016/j.knsys.2022.109609](https://doi.org/10.1016/j.knsys.2022.109609) (2022).

6. Baruch, E. B. & Keller, Y. Joint detection and matching of feature points in multimodal images. *IEEE Transactions on Pattern Analysis Mach. Intell.* **44**, 6585–6593, DOI: [10.1109/TPAMI.2021.3092289](https://doi.org/10.1109/TPAMI.2021.3092289) (2022).
7. Jiang, B., Sun, P. & Luo, B. Glnet: graph learning-matching convolutional networks for feature matching. *Pattern Recognit.* **121**, 108167, DOI: [10.1016/j.patcog.2021.108167](https://doi.org/10.1016/j.patcog.2021.108167) (2022).
8. Chen, Q. *et al.* Horticultural image feature matching algorithm based on improved orb and lk optical flow. *Remote. Sens.* **14**, 4465, DOI: [10.3390/rs14184465](https://doi.org/10.3390/rs14184465) (2022).
9. Gong, D., Huang, X., Zhang, J., Yao, Y. & Han, Y. Efficient and robust feature matching for high-resolution satellite stereos. *Remote. Sens.* **14**, 5617, DOI: [10.3390/rs14215617](https://doi.org/10.3390/rs14215617) (2022).
10. Gong, X. *et al.* Feature matching for remote-sensing image registration via neighborhood topological and affine consistency. *Remote. Sens.* **14**, 2606, DOI: [10.3390/rs14112606](https://doi.org/10.3390/rs14112606) (2022).
11. DeTone, D., Malisiewicz, T. & Rabinovich, A. Superpoint: Self-supervised interest point detection and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 224–236 (2018).
12. Sarlin, P.-E., DeTone, D., Malisiewicz, T. & Rabinovich, A. Superglue: Learning feature matching with graph neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4938–4947 (2020).
13. Gao, Y. & Zhao, L. Coarse trvo: a robust visual odometry with detector-free local feature. *J. Adv. Comput. Intell. Intell. Informatics* **26**, 731–739, DOI: [10.20965/jaciii.2022.p0731](https://doi.org/10.20965/jaciii.2022.p0731) (2022).
14. Kim, D., Lee, J., Cho, M. & Kwak, S. Detector-free weakly supervised group activity recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 20051–20061, DOI: [10.1109/CVPR52688.2022.01945](https://doi.org/10.1109/CVPR52688.2022.01945) (2022).
15. Mok, T. C. W. & Chung, A. C. S. Affine medical image registration with coarse-to-fine vision transformer. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 20803–20812, DOI: [10.1109/CVPR52688.2022.02017](https://doi.org/10.1109/CVPR52688.2022.02017) (2022).
16. Fan, Y., Wang, F. & Wang, H. A transformer-based coarse to-fine wide-swath sar image registration method under weak texture conditions. *Remote. Sens.* **14**, 1175, DOI: [10.3390/rs14051175](https://doi.org/10.3390/rs14051175) (2022).
17. Chen, J. *et al.* Trans morph: transformer for unsupervised medical image registration. *Med. Image Analysis* **82**, 102615, DOI: [10.1016/j.media.2022.102615](https://doi.org/10.1016/j.media.2022.102615) (2022).
18. Ma, M., Xu, Y., Song, L. & Liu, G. Symmetric transformer-based network for unsupervised image registration. *Knowledge-Based Syst.* **257**, 109959, DOI: [10.1016/j.knosys.2022.109959](https://doi.org/10.1016/j.knosys.2022.109959) (2022).
19. Sun, J., Shen, Z., Wang, Y. *et al.* Loftr: Detector-free local feature matching with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8922–8931 (2021).
20. Giang, K. T., Song, S. & Jo, S. Topicfm: Robust and interpretable topic-assisted feature matching. In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, 2447–2455 (2023).
21. Truong Giang, K., Song, S. & Jo, S. Topicfm+: Boosting accuracy and efficiency of topic-assisted feature matching. *arXiv e-prints* (2023). ArXiv:2307.00485.
22. Zhou, Q., Sattler, T. & Leal-Taixe, L. Patch2pix: Epipolar-guided pixel-level correspondences. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4669–4678 (2021).
23. Huang, D., Chen, Y., Liu, Y. *et al.* Adaptive assignment for geometry aware local feature matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5425–5434 (2023).
24. Chen, H., Luo, Z., Zhou, L. *et al.* Aspanformer: Detector-free image matching with adaptive span transformer. In *European Conference on Computer Vision*, 20–36 (Springer Nature Switzerland, 2022).
25. Wang, Q., Zhang, J., Yang, K. *et al.* Matchformer: Interleaving attention in transformers for feature matching. In *Proceedings of the Asian Conference on Computer Vision*, 2746–2762 (2022).
26. Edstedt, J., Athanasiadis, I., Wadenbäck, M. *et al.* Dkm: Dense kernelized feature matching for geometry estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17765–17775 (2023).
27. Edstedt, J., Sun, Q., Bökman, G. *et al.* Roma: Revisiting robust losses for dense feature matching. *arXiv preprint arXiv:2305.15404* (2023).
28. Truong, P., Danelljan, M., Van Gool, L. *et al.* Learning accurate dense correspondences and when to trust them. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5714–5724 (2021).

29. Szegedy, C., Liu, W., Jia, Y. *et al.* Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1–9 (2015).
30. Müller, M. M. & Kornblith, S. An introduction to distance metric learning. *arXiv preprint arXiv:2006.10069* (2020).
31. Muja, M. & Lowe, D. G. Fast approximate nearest neighbors with automatic algorithm configuration. In *International conference on computer vision theory and applications*, vol. 2, 331–340 (2009).
32. Zheng, Y., Yu, Y. & Fang, C. A deep learning approach to unsupervised image quality assessment. *IEEE Transactions on Cybern.* **49**, 1182–1195 (2018).
33. Muja, M. & Lowe, D. G. Fast approximate nearest neighbors with automatic algorithm configuration. In *International conference on computer vision theory and applications*, vol. 2, 331–340 (2009).
34. Lowe, D. G. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.* **60**, 91–110 (2004).
35. Di, Y. *et al.* Femip: detector-free feature matching for multimodal images with policy gradient. *Appl. Intell.* DOI: [10.1007/s10489-023-04659-5](https://doi.org/10.1007/s10489-023-04659-5) (2023).
36. Liu, C., Yuen, J. & Torralba, A. Sift flow: Dense correspondence across scenes and its applications. *IEEE Transactions on Pattern Analysis Mach. Intell.* **33**, 978–994 (2010).
37. Rublee, E., Rabaud, V., Konolige, K. & Bradski, G. Orb: An efficient alternative to sift or surf. In *2011 International Conference on Computer Vision*, 2564–2571 (IEEE, 2011).
38. Bay, H., Tuytelaars, T. & Van Gool, L. Surf: Speeded up robust features. In *European Conference on Computer Vision*, 404–417 (Springer, Berlin, Heidelberg, 2006).
39. Tyszkiewicz, M., Fua, P. & Trulls, E. Disk: Learning local features with policy gradient. In *Advances in Neural Information Processing Systems*, vol. 33, 14254–14265 (2020).
40. Wang, Q., Zhou, X., Hariharan, B. & Snavely, N. Learning feature descriptors using camera pose supervision. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, 757–774 (Springer, 2020).
41. Tang, S., Zhang, J., Zhu, S. & Tan, P. Quadtree attention for vision transformers. *arXiv preprint arXiv:2201.02767* (2022).
42. Yu, J. *et al.* Adaptive spot-guided transformer for consistent local feature matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 21898–21908 (2023).
43. Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K. & Yuille, A. L. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Transactions on Pattern Analysis Mach. Intell. (TPAMI)* **40**, 834–848 (2018).
44. Badrinarayanan, V., Kendall, A. & Cipolla, R. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Transactions on Pattern Analysis Mach. Intell. (TPAMI)* **39**, 2481–2495 (2017).
45. Zhao, H., Shi, J., Qi, X., Wang, X. & Wang, L. Pyramid scene parsing network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2881–2890 (2017).
46. LeCun, Y., Bengio, Y. & Hinton, G. Gradient-based learning applied to document recognition. *Proc. IEEE* **86**, 2278–2324 (1998).
47. Vaswani, A. *et al.* Attention is all you need. In *Advances in neural information processing systems*, 5998–6008 (2017).
48. Ronneberger, O., Fischer, P. & Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, 234–241 (Springer, Cham, 2015).
49. Loshchilov, I. & Hutter, F. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101* (2017).
50. Loshchilov, I. & Hutter, F. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983* (2016).
51. Mishchuk, A., Mishkin, D., Radenovic, F. & Matas, J. Working hard to know your neighbor’s margins: Local descriptor learning loss. In *Advances in Neural Information Processing Systems*, vol. 30 (2017).
52. Balntas, V., Riba, D. E. & Mikolajczyk, K. Learning local feature descriptors with triplets and shallow convolutional neural networks. In *Proceedings of the British Machine Vision Conference (BMVC)*, 119.1–119.11 (BMVA Press, 2016). Available from: <https://dx.doi.org/10.5244/C.30.119>.
53. Liao, Y. *et al.* Feature matching and position matching between optical and sar with local deep feature descriptor. *IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens.* **15**, 448–462, DOI: [10.1109/JSTARS.2021.3134676](https://doi.org/10.1109/JSTARS.2021.3134676) (2022).
54. Han, X., Leung, T., Jia, Y., Sukthankar, R. & Berg, A. C. Matchnet: Unifying feature and metric learning for patch-based matching. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3279–3286 (2015).

Author contributions statement

R.T. conceived the method, wrote the manuscript, and designed the experiment. Y.L. improved the methods, checked and revised the manuscript, and provided hardware support. W.W. contributed to the theoretical analysis and algorithm development. Q.D. conducted related investigations. J.L. prepared the data. W.J. performed the experimental analysis. Yp.L. contributed to the experimental setup and assisted with data analysis. X.Q. performed data preprocessing. All authors reviewed the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to Y.L.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.