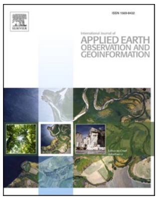
International Journal of Applied Earth Observation and Geoinformation 125 (2023) 103574

Contents lists available at ScienceDirect

International Journal of Applied Earth Observation and Geoinformation

journal homepage: www.elsevier.com/locate/jag



Coarse-to-fine matching via cross fusion of satellite images

Liangzhi Li [a,b], Ling Han [c,*], Kyle Gao [d], Hongjie He [b], Lanying Wang [b], Jonathan Li [b,d,*]

[a] *College of Geological Engineering and Geomatics, Chang'an University, Xi'an, SX 710064, China*
[b] *Department of Geography and Environmental Management, University of Waterloo, Waterloo, ON N2L 3G1, Canada*
[c] *School of Land Engineering, Chang'an University, Xi'an, SX 710064, China*
[d] *Department of Systems Design Engineering, University of Waterloo, Waterloo, ON N2L 3G1, Canada*

ARTICLE INFO

Keywords:
Coarse-to-fine matching
Cross fusion
Satellite images
Self-attention
Registration

ABSTRACT

The registration of multimodal satellite images is essential for a prerequisite for accruing complementary observational data. Nevertheless, the differential imaging nuances amongst non-linear radiometric multimodal images precipitate a complexity in keypoint detection, rendering it a great challenge. This complexity exacerbates the difficulty encountered in matching multimodal satellite images. In this paper, a dual-branch cross fusion network (DF-Net) is proposed for the purpose of satellite image registration. DF-Net relies on the self-attention granted to a pair of images, thereby providing cross-modal fusion feature descriptions. Initially, reference and sensed images are deployed as inputs for the dual-branch network, which in turn engenders feature descriptions of both high and low resolution, respectively. Sequentially, the matching of individual feature descriptions is anchored on the low-resolution feature map, paving the way for the establishment of coarse matching correspondences. Subsequently, the outcomes of these coarse correspondences are transposed onto the feature map with a higher resolution, thereby generating fine matching results for each coarse correspondence. An exhaustive set of qualitative and quantitative assessments have been administered on three satellite image datasets encompassing a diverse range of scenarios. The average Repeatability (Rep.), Mean Matching Accuracy (MMA), and Root-Mean-Square Error (RMSE) of the DF-Net applied to three large-scale satellite images were recorded to be 0.71, 0.65, and 2.34, respectively. These findings buttress the proficiency of the proposed strategy in facilitating cross-modal matching and bear testimony to the sterling performance of the method proposed.

1. Introduction

Registration is the basis for multi-source satellite information representation between images of the same scene, which is a prerequisite for band fusion, change detection, and image stitching (Ma et al., 2022). Geographic alignment of satellite images can eliminate large geometric errors. However, the correspondence between multi-source satellite images still has tens of pixels of error due to the difference in imaging perspectives. Consequently, relying solely on geographic alignment to attain sub-pixel matching for satellite images is infeasible, necessitating the implementation of a matching process (Zhang et al., 2023). Nonetheless, the presence of nonlinear radiometric disparities between multimodal satellite images and local geometric distortions renders the matching process notably challenging (Fu et al., 2020; Deng et al., 2023).

Most existing satellite image matching methods incorporate the “detection-description-matching-geometry constraint” step. There are two main categories of image matching techniques: feature-based and area-based. These methods are differentiated by the way in which keypoints are generated (Haskins et al., 2020; Jiang et al., 2019). The feature-based methods facilitate correspondence identification through the application of similarity metrics for measuring feature resemblances (Wu et al., 2022). The area-based methods solve the alignment problem by finding matches directly from the whole image or patch images using similarity metrics. These methods are sufficient for most homologous satellite image matching. However, these feature detectors may not obtain keypoints with repeatability when considering two different modalities in the same scene due to the satellite sensor’s shooting angle, multi-temporal features, intensity, and variations in viewing angle. Even the best feature descriptor and matching strategy cannot find a match between two images without repeated keypoints. Fig. 1 shows the key points (blue) in SAR and optical satellite images, respectively, obtained by the Harris corner detector.

For some area-based matching methods based on similarity metrics,

[*] Corresponding authors at: Department of Geography and Environmental Management, University of Waterloo, Waterloo, ON N2L 3G1, Canada (J. Li).
E-mail addresses: hanlin@chd.edu.cn (L. Han), junli@uwaterloo.ca (J. Li).

https://doi.org/10.1016/j.jag.2023.103574
Received 9 June 2023; Received in revised form 22 October 2023; Accepted 14 November 2023
Available online 17 November 2023

they can be employed in a pixel-by-pixel search to remedy this problem. Matches with high similarity scores are used as the final correspondence from the results of these searches, avoiding the step of feature detection. Nonetheless, this approach proves to be time-consuming and necessitates considerable computational resources, thereby constraining its applicability to satellite images. Certain studies employed deep leaning networks to map the positions of patches on reference images (Li et al., 2021, 2023). While these techniques can enhance the matching efficiency by avoiding pixel-by-pixel searches, they do not consider cross-modal feature similarities, which can be heavily influenced by the dataset used. As a result, the accuracy of these methods may be limited in certain circumstances.

Recently, several works obtained correspondences between two natural images by establishing matching at the pixel level (e.g., Jiang et al., 2022; Revaud et al., 2019; Sun et al., 2021). As these methods are primarily designed for natural image matching, they need to be adapted to accommodate multi-source remote sensing images. For instance, features extracted from satellite images using the convolutional neural networks (CNNs) without information interaction may yield dissimilar feature descriptions (Li et al., 2022; Liu et al., 2023). Moreover, CNNs possess a restricted receptive field, potentially resulting in indistinguishable feature descriptions.

Therefore, establishing matching between satellite images requires overcoming the following problems.

- The first problem is that of the receptive field for feature detection. Since CNNs have a restricted receptive field, they lack a wider variety of feature fusion and may not distinguish inconspicuous feature descriptions.
- The second issue is the time-consuming nature of cost volume searching by the pixel-by-pixel search method. Establishing matching correspondence for each pixel in a given satellite image is time-consuming.
- The third issue is that of cross-modal feature similarity. This refers to the challenge of extracting features from satellite images using a shared network weight, which may not adequately capture the similarities necessary for successful cross-modal matching.

To mitigate these limitations, we introduce a dual-branch fusion network, denoted as DF-Net, designed to identify correspondences between reference and sensed images by employing a coarse-to-fine matching strategy. Our main contributions include:

- We construct a dual-branch network with self-attention for providing a larger range of receptive field to obtain feature descriptions with global dependencies.
- To increase the efficiency, we utilize a coarse-to-fine matching strategy. This involves initially extracting coarse matches at low resolution and then selecting the matching results with high confidence for fine matching at the sub-pixel level.
- To acquire feature descriptions exhibiting cross-modal similarity, we propose an interactive fusion module that generates feature descriptions contingent upon both images.

The rest of the paper is organized as follows. Section 2 provides a comprehensive literature review of satellite image matching and self-attention. Section 3 details our proposed method with an emphasis on its functional components. Section 4 presents and discusses our experimental results. Section 5 concludes the paper.

2. Related studies

2.1. Satellite image matching

Feature-based matching methods: Numerous research endeavors have focused on enhancing the scale-invariant feature transform (SIFT) (Ng and Steven, 2003) method for multimodal satellite image feature detection, encompassing variants such as SAR-SIFT (Wang et al., 2021). Although these SIFT variations effectively create local features with distinguishability, their performance may be suboptimal in alternative application scenarios. The interest in multimodal image matching tasks has surged in recent years, culminating in the development of various algorithms, such as the partial intensity invariant feature descriptors (PIIFD) (Chen et al., 2010), Distinctive Order Based Self Similarity descriptor (DOBSS) (Sedaghat and Mohammadi, 2019), and Modified RIFT (Chen et al., 2022). These multimodal image matching techniques have exhibited success within the realm of multimodal satellite images (Meng et al., 2021).

The matching performance of multimodal satellite images with nonlinear radiometric differences can be significantly improved by a deep learning method (Wang et al., 2022). Yang et al. (2018) introduced a multiscale feature description founded on trained CNNs to enhance robustness, thereby improving registration robustness through the gradual increase of inlier selection. Ye et al. (2018) combined SIFT and CNN features information fusion for satellite image matching, which provided a high-level information for satellite image registration. Since these methods introduce deep learning methods on multimodal satellite

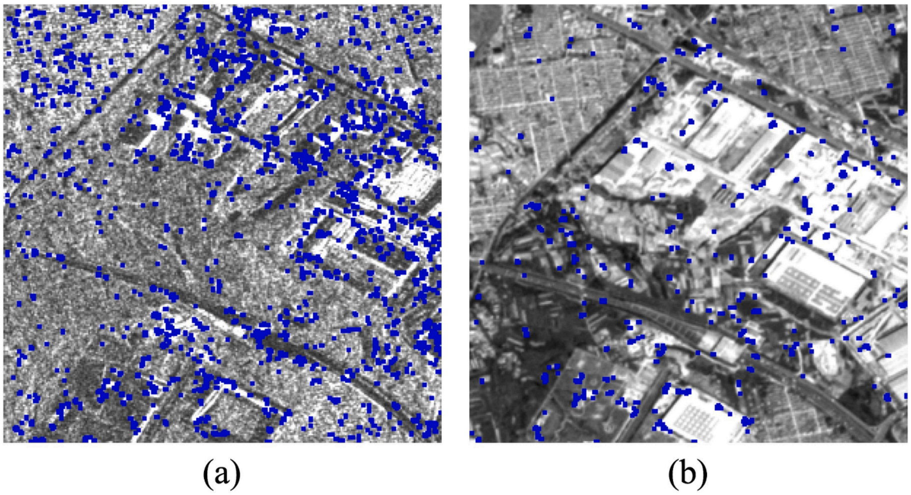


Fig. 1. Key points (blue) extracted by the Harris corner detector from (a) SAR, and (b) optical satellite imagery. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

image feature descriptions, they are still based on the description of salient features.

Area-based matching methods: Area-based techniques are prevalent approaches for cross-modal matching. These methods identify feature points within two images and subsequently establish correspondences within a local search window. Correspondence measurement in the search window is determined by the similarity among local features. Widely used similarity metrics encompass the normalized cross-correlation (NCC) (Cui et al., 2020), Sum of Squared Differences (SSD) (Zhang et al., 2020), mutual information (MI) (Sengupta et al., 2022), channel features of orientated gradients (CFOG) (Ye et al., 2019), and Automatic Registration of Remote-Sensing Images (ARRSI) (Wong et al., 2007). While SSD and NCC exhibit sensitivity to nonlinear radiometric differences, MI and CFOG demonstrate increased robustness against radiometric disparities. Nevertheless, MI, CFOG, and ARRSI are ill-suited for multimodal satellite image alignment due to their vulnerability to template distortion.

To counteract the limitations, some methods have been developed to generate similarities based on local features (Yao et al., 2022; Gao et al., 2022; Quan et al., 2022). However, some methods employ weight sharing strategies to derive similar features through fully connected layer stacking. Although these approaches endeavor to accommodate similarity across the entire dataset, their effectiveness may be heavily contingent upon the availability of large and diverse datasets. Furthermore, the final matches are generally acquired via pixel-level searches, which can impose substantial computational demands.

2.2. Self-attention

Recently, Transformer and self-attention structures have transcended the boundaries of natural language processing and become popular research directions in computer vision. In the specific area of image matching, their potential has been widely explored and confirmed. Chen et al. (2022) proposed a Transformer-based detector less matching method. This method employs an adaptive attention span to adjust the attention, which ensures long-distance dependence, while reinforcing the fine-grained attention between highly correlated pixels in the matching task. Meanwhile, Lu et al. (2023) proposed a parallel attention mechanism (ParaFormer). By integrating self-attention and cross-attention, this approach not only improves the accuracy of matching, but also significantly improves the computational efficiency. In cross-view geo-localization, Tian et al. (2022) first used semantic segmentation to distinguish different image regions, and then used Transformer to explicitly utilize the properties of self-attention Perform Matching.

Local Feature Matching with Transformers (LoFTR) (Sun et al., 2021) utilizes self-attention in the Transformer to map feature descriptors in two images. Wang et al. (2022) proposed a hierarchical extraction and matching transformer called MatchFormer, which intertwines self-attention for feature extraction and cross-attention for feature matching to achieve efficiency, robustness, and accuracy with state-of-the-art results in four different benchmarks. On cross-modal matching, both Transformer and self-attention demonstrate their strong potential. With these studies, it is reasonable to believe that these methods will be widely used in more applications in the future.

2.3. Multimodal feature fusion

In recent years, research on cross-modal neural networks has made tremendous progress (Xie et al., 2023). Wang et al. (2015) was one of the early researchers in the field of cross-modal, and they proposed a clustering-sensitive cross-modal relevance learning framework to address the challenges of processing large-scale Web data. Wei et al. (2016) further demonstrated the benefits of CNN visual features for cross modal retrieval with their extensive experiments on five popular publicly available datasets. While He et al. (2016) proposed a new architecture, their results clearly showed that the architecture can efficiently learn representations with good semantics to achieve superior cross-modal retrieval performance9.

Cangea et al. (2019) improved upon multimodal deep learning by proposing a new cross-modal approach that extends the previous cross connectivity that only transfers information between processing-compatible data streams. Subsequently, Khowaja and Lee (2020) proposed that hybrid fusion has different representations than the basic modality, which provides a new direction for cross-modal learning streams to be training with new directions.

Xu et al. (2020) research further deepened the understanding in this area by demonstrating the advantages of their P3S approach by comparing it with 15 state-of-the-art methods on four widely used cross-modal datasets. Liu et al. (2020) conducted extensive experiments on three benchmark datasets, which demonstrated that their model is at the state-of-the-art in cross modal retrieval to state-of-the-art results. Wei and Zhou (2021) pointed out the technical challenges of cross-modal communication in co-transmitting and processing audio, visual, and haptic signals, but also emphasized the potential for AI technologies to support this. Geigle et al. (2022) also conducted experiments on a variety of cross-modal retrieval benchmarks, and their approach compares favorably to state-of-the-art in terms of accuracy and efficiency. cross-coders in terms of both accuracy and efficiency. Prakash et al. (2021) proposed two-channel fusion of images and point clouds for self-attention to achieve end-to-end autopilot, reducing collisions by 76 % compared to geometry-based fusion.

These above approaches, constructed for different cross-modal information fusion methods in different tasks, have made a lot of progress in the existing research. However, in multimodal remote sensing image matching, instead of establishing the fusion of information in two images, the similarity of key features is established based on two images as conditions. Therefore, based on the two-channel multimodal information fusion. we propose an attention structure with cross-modality to learn the similarity of feature representations of two images.

3. Method

In this study, we present a dual-branch fusion network (DF-Net) that leverages self-attention to enhance the global receptive field and facilitate cross-modal information fusion for satellite image registration. In the subsequent section, we present a comprehensive exposition of DF-Net, encompassing an overarching overview, the dual-branch network, self-attention modules, and loss functions.

3.1. Framework of proposed DF-Net

DF-Net, as depicted in Fig. 2, is an innovative framework designed to determine the correspondence between image pairs by employing a systematic coarse-to-fine strategy. At its core, this methodology is based on a dual-branch network system complemented by a self-attention module, which collectively facilitates precise image matching.

Feature Extraction in DF-Net: The primary focus in the DF-Net framework is the extraction of relevant features from input images, which is accomplished in two pivotal steps:

1. Dual-Branch Network Structure: The essence of this structure is rooted in residual networks, specifically orchestrated in Stages 2 and 4. The outcome of this process yields two distinct maps: the high-resolution map, scaled at $\frac{1}{2}$ of the input size, and the low-resolution map, scaled at $\frac{1}{8}$ of the input size.
2. Self-Attention Mechanism: Perpendicular to the aforementioned dual-branch network, the self-attention module plays a crucial role. It addresses the residual block's receptive field limitations by extending its range. More importantly, it fosters an environment for information cross-fusion, enabling both branches of the network to

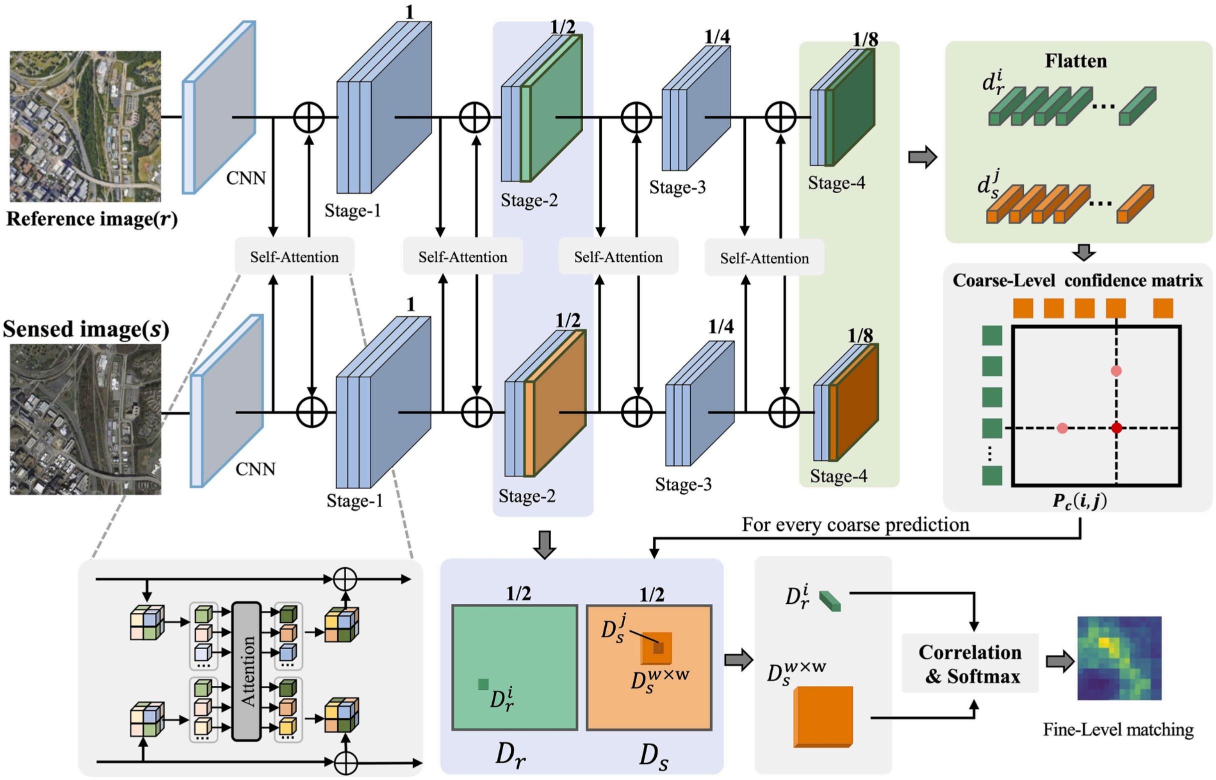


Fig. 2. Framework of the proposed DF-Net. A dual-branch residual networks produce high- and low-resolution maps at 1/2 and 1/8 input sizes. Self-attention is integrated orthogonally. Initial coarse matching uses one-dimensional vectors from low-resolution maps to form a confidence matrix. This coarse matrix then guides fine matching at 1/2 input size within a w × w local window.

harness features from each other. This cross-modal interaction ensures that the output feature descriptions encapsulate cross-modal similarities.

DF-Net Matching Process: Post feature extraction, the DF-Net framework delves into the matching phase, which is also bifurcated into two steps:

- Step 1, Coarse Matching: Here, the low-resolution feature maps derived from Stage 4, which are scaled to $\frac{1}{8}$ of the original image size, are linearly transformed into one-dimensional vectors, denoted as d_r^i and d_s^j. These vectors are subsequently processed through a unique matching layer, generating a coarse-level confidence matrix $P_c(i,j)$. Utilizing a predefined confidence threshold, selections from $C_m(i,j)$ are made, culminating in the coarse matching prediction, M_c.
- Step 2, Fine Matching: Building on the results of the coarse matching, the positions $P_c(i,j)$ that belong to M_c are projected onto the high-resolution feature map from Stage 2. Here, D_r^i epitomizes the central feature of the coarse matches, while $D_s^{w\times w}$ is a subset cropped from Stage 2, characterized by its $w\times w$ dimensions. The final step entails refining the matches within this window to achieve sub-pixel matching precision.

ResNet-50 (He et al., 2016) was used as the underlying structure of a dual-branch network for extracting features from the two images, serves as the foundational structure of the dual-branch network. This choice ensures the robust extraction of features from both images in the pair. Moreover, the downsampling procedures within the residual network block were re-envisioned to align with the goals of DF-Net. As a result, the output features from Stages 1 through 4 manifest as $\frac{1}{1}, \frac{1}{2}, \frac{1}{4}$, and $\frac{1}{8}$ of the original image size, respectively.

3.2. Two-branch residual networks

DF-Net is comprised of parallel residual blocks, with residual networks featuring jump connections in each block, facilitating training and optimization. ResNet-50 serves as the primary backbone of the proposed method. Two images are separately fed in the branch network, where feature extraction is performed via a convolution operation and a four-stage residual network (Stages 1, 2, 3, 4). The strides and feature maps in each residual block are as such; stride = 1 in Stage 1, while Stages 2, 3, 4 = 2, to obtain maps with sizes of 1/2, 1/4, 1/8, respectively. The number of feature channels for each stage is adjusted to 64, 128, 256 and 512, respectively, to improve the efficiency of matching. Finally, the feature maps of 1/2 and 1/8 original image size are selected for coarse to fine matching.

3.3. Vertical cross-attention module

The vertical cross-attention module is founded on the self-attention mechanism. Image-based SA enables the capture of global information by enlarging the receptive field, facilitating the internal correlation, and subsequently reducing dependency on external information. We employed self-attention to bolster information interaction between local features and the global sequence. Furthermore, we derived similarity feature descriptions with cross-modality by enhancing the correlation through self-attention. In contrast to the local feature extraction of CNN operations, stacked convolutional layers are utilized to extend the receptive field. However, this approach may impose additional computational overhead. SA offers an efficient means of modeling global contextual information using key, query, and value components. This process enables the vertical self-attention module to effectively capture global contextual information while maintaining a manageable computational burden.

Fig. 3 depicts the cross-attention module perpendicular to the two-

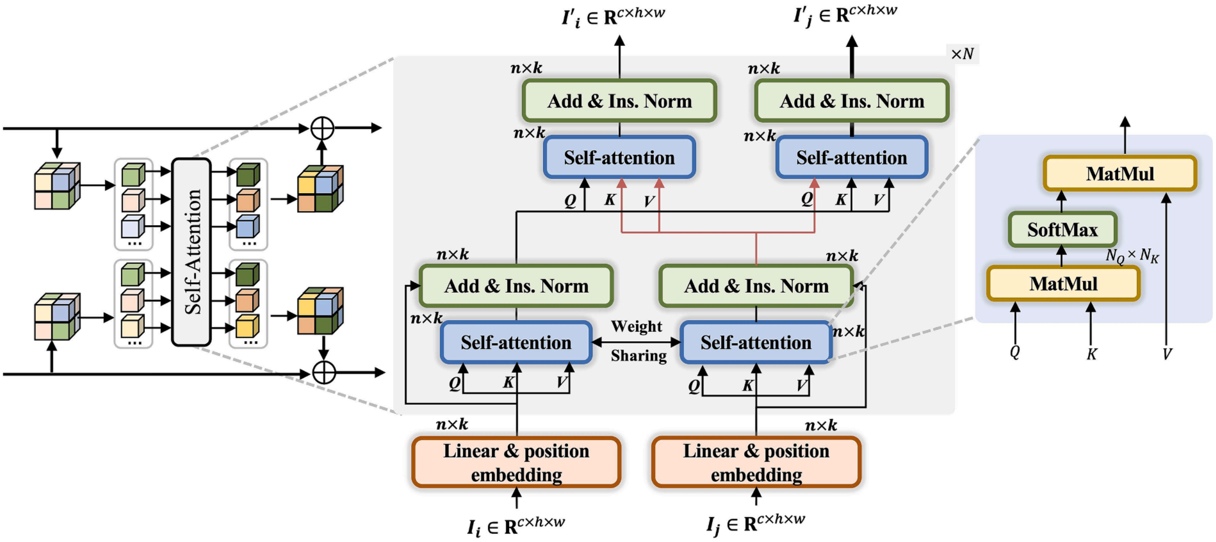

$I'_i \in \mathbf{R}^{c\times h\times w}$
$I'_j \in \mathbf{R}^{c\times h\times w}$
×N
n×k
Add & Ins. Norm
Self-attention
Q
K
V
Weight
Sharing
Linear & position embedding
$I_i \in \mathbf{R}^{c\times h\times w}$
$I_j \in \mathbf{R}^{c\times h\times w}$
Self-Attention
MatMul
SoftMax
$N_Q \times N_K$

Fig. 3. Vertical cross-attention structure. Following linear projection and position encoding, the outputs of each stage within two branches, $\mathrm{I_i} \in \mathbf{R}^{\mathrm{c\times h\times w}}, \mathrm{I_j} \in \mathbf{R}^{\mathrm{c\times h\times w}}$, are input to the weight-sharing self-attention for increasing the receptive field of feature extraction. Outputs are employed as inputs for the interactive fusion of information via feature summation (Add), linear projection (Ins), and normalization (Norm) operations. In the feature fusion stage, the output K, V is exchanged and the feature feature similarity $\left(\mathrm{I}_\mathrm{i}^{'} \in \mathbf{R}^{\mathrm{c\times h\times w}}, \mathrm{I}_\mathrm{j}^{'} \in \mathbf{R}^{\mathrm{c\times h\times w}}\right)$ with cross-modality is obtained after information exchange by self-attention. N denotes the number of network layer repetitions.

branch network. The features obtained from each stage in the two branches are represented as F_r and F_s via patching embedding, linear transformation, and position embedding, respectively. F_r and F_s are then input to the shared weight module consisting of a fully connected layer (Long et al., 2015) and self-attention. The local features in F_r and F_s query the global contextual features by self-attention, thus, the two feature maps obtain a larger receptive field in their spatial ranges.

During the feature cross-fusion phase, reference features (F_r) and sensed features (F_s) are associated with respective query (Q), key (K), and value (V) elements. similarity information is obtained following the cross-fusion process via cross-attention. Conceptually, the SA interaction fusion operation identifies pertinent information by evaluating the resemblance between feature descriptions in input images. This cross-fusion operation is performed four times over the course of the feature extraction procedure. In every iteration, F_r and F_s exchange K and V components as input while utilizing Q to extract attention from the K and V elements. Through the assessment of similarity between reference and sensed images, the cross-fusion operation efficiently supports feature extraction and bolsters the model's ability to capture significant information from multimodal images with enhanced efficacy.

Fig. 3 shows the cross-fusion mechanism. Following linear projection and position encoding, the outputs of each stage within two branches, $I_i \in \mathbf{R}^{c\times h\times w}$, $I_j \in \mathbf{R}^{c\times h\times w}$, are input to the weight-sharing SA for increasing the receptive field of feature extraction. Outputs are employed as inputs for the interactive fusion of information via feature summation (Add), linear projection (Ins), and normalization (Norm) operations. In the feature fusion stage, the output K, V is exchanged and the feature feature similarity ($I_i^{'} \in \mathbf{R}^{c\times h\times w}$, $I_j^{'} \in \mathbf{R}^{c\times h\times w}$) with cross-modality is obtained after information exchange by self-attention. N denotes the number of network layer repetitions. Given input tensors I^1, I^2, $I^3 \in \mathbf{R}^{d_{l\times 1}}$, each of these tensors symbolizes distinct feature descriptions or representations obtained from the input data, where $I^i, i \in 1, 2, 3$ are transformed by $W^q \in \mathbf{R}^{d_k\times d_l}$, $W^k \in \mathbf{R}^{d_k\times d_l}$, $W^v \in \mathbf{R}^{d_l\times d_l}$ to obtain $q^i \in \mathbf{R}^{d_k\times 1}$, $k^i \in \mathbf{R}^{d_k\times 1}$, and $v^i \in \mathbf{R}^{d_l\times 1}$. The matrix is $A = (I^1, I^2, I^3) \in \mathbf{R}^{d_{l\times 3}}$, then Q, K, V

$$\begin{cases} Q = W^q \cdot A \\ K = W^k \cdot A \\ V = W^v \cdot A \end{cases} \tag{1}$$

The output matrix is $T = V\cdot\mathrm{softmax}\left(\frac{K^\top \cdot Q}{\sqrt{d^k}}\right)$. The final output $t^1, t^2, t^3, \cdots, t^i$ is transformed by the fully connected layer and the BN layer. Fig. 4 shows an example of the feature extraction for receptive fields and cross-modal similarity matching, where self-attention mechanism is utilized to broaden the receptive field by obtaining correlations between each feature description via a long-range dependency mechanism.

3.4. Coarse matching

For the 1/8 low-resolution feature maps, they are differentiable by the matching layer after conversion by calculating the score matrix between features. The score matrix S is calculated by $S(i,j) = \frac{1}{\tau}\cdot\langle d_r(i), d_s(j)\rangle$. Dual-softmax, is applied in both dimensions of $S(i,j)$ to obtain the matching confidence of mutual nearest neighbors. The detailed formula for calculating the matching confidence is presented below.

$$P_c(i,j) = \mathrm{softmax}(S(i,\cdot))_j \cdot \mathrm{softmax}(S(\cdot,j))_i \tag{2}$$

$$M_c = \{(i,j) | \forall (i,j) \in \mathrm{MNN}(P_c), P_c(i,j) \geq \theta_c\} \tag{3}$$

Based on the matching confidence, we apply the mutual nearest neighbor (MNN) criterion to filter the possible coarse matches. A suitable threshold is chosen to identify the coarse matches that will be input for subsequent fine matching. The prediction for coarse matching can be expressed as follows: The prediction for coarse matching can be formulated as follows:

where the term θ_c is a predetermined threshold value, employed to sieve out matches with confidence levels below this threshold.

3.5. Fine matching

Once coarse matches are established, these correspondences undergo refinement to the original image size via fine matching. To obtain further refined matching results on coarse matches, we propose a fully connected layer-based similarity method. For each coarse match M_c,

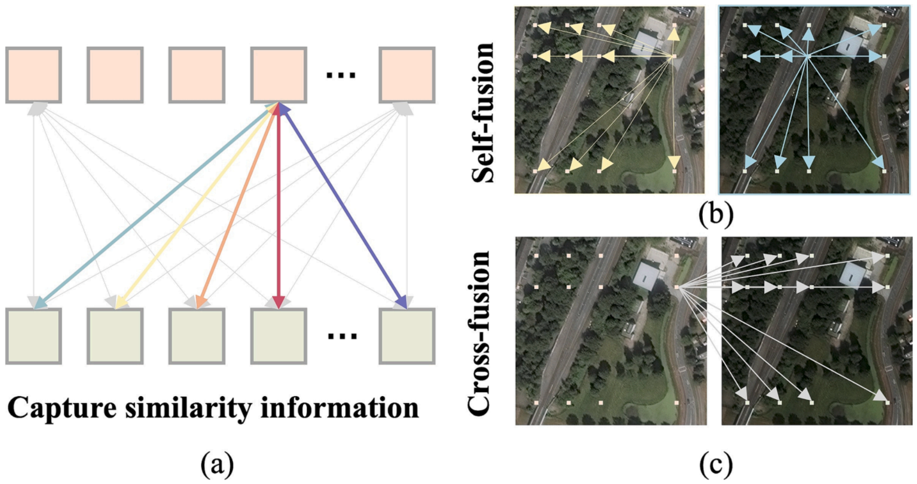


Fig. 4. Illustration of the receptive field and cross-fusion. (a) Information fusion for capturing similarity information, (b) and (c) Fusion processes on satellite images, respectively.

their positions (i,j) are first positioned on a feature map of size $1/2$. Then, two matching windows of size $w \times w$ are cropped separately. We correlate the center vector of D_r^i with all vectors of $D_r^{w\times w}$ and input them into the proposed fully connected layer-based similarity network to map the similarity to the interval $[0,1]$, and the detailed procedure is as follows:

$$S(i,j) = Sigmod\left(Dense\left(R_i - S_j\right)^2\right) \tag{4}$$

The heat map results represent the center vector of D_r^i with all vectors matching responses of $D^{w\times w}r$, with the maximum value indicating the result, where *Dense* refers to fully connected layers.

3.6. Loss function

The combined loss function incorporates both coarse and fine matching loss functions, as illustrated below.

Coarse matching loss function. For the confidence matrix $(P_c(i,j))$ returned by coarse matching, the coarse matching probability loss is given by a negative logarithm. The labels of the true confidence matrix during training are calculated based on the projection transformation matrix of the two images. Employing the method proposed by SuperGlue (Liu et al. 2021), the ground truth for coarse matching, denoted as M_c^{gt}, is determined by identifying the mutual nearest neighbors within the two sets of low-resolution feature map grids.

The distance between the low-resolution feature map grids is quantified using the reprojection distance of their central locations. By minimizing the negative log-likelihood of the grids in M_c^{gt}, we effectively reduce this distance.

$$L_1 = -\frac{1}{\left|M_c^{gt}\right|}\sum_{(i,j)\in M_c^{gt}} \log P_c(i,j) \tag{5}$$

Fine-matching loss function. Root-mean-square error loss is used for fine-matching optimization. The center vector D_r^i produces only one correspondence with all vectors $D_r^{w\times w}$, and the matching is transformed into a regression problem using the root mean square error. j_{gt} is computed by the projection transformation matrix of i from D_r to D_s. For L_2, if the distorted position of i is not within the local window $D_r^{w\times w}$, we will ignore $\left(i, j_{gt}\right)$ and the gradient will not be backpropagated during the training period.

$$L_2 = \sqrt{\frac{1}{M}\sum_{i=1}^{M}\left(x_i - x_{gt}\right)^2 + \left(y_i - y_{gt}\right)^2} \tag{6}$$

where x_i, y_i denote the matching positions of i on the x,y-axes. x_{gt}, y_{gt} denote the ground truth values.

4. Results and discussion

4.1. Evaluation metrics

The network's matching performance is evaluated based on the following assessment protocol:

Repeatability (Rep.): The repetition rate is the ratio of the number of pixels with the same position between two images at the same threshold that are detected to the total number of detections. The detection performance of the network is evaluated using the repeatability (n/N), where n, N are the number of repeatable and all responses obtained, respectively.

Mean Matching Accuracy (MMA): The mean proportion of accurate correspondences within an image at a designated pixel threshold.

Root-Mean-Square Error (RMSE): RMSE is utilized to evaluate the comprehensive performance.

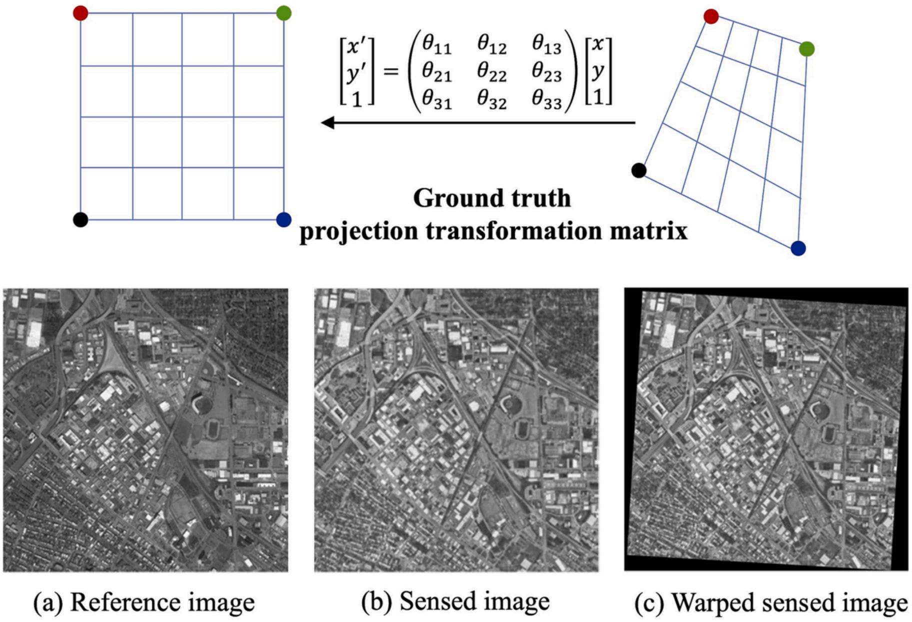
$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left[\left(\Delta_i\right)^2 + \left(\Delta_j\right)^2\right]} \tag{7}$$

where N is the total matching points, Δ_i and Δ_j are the matching distances, respectively.

4.2. Database and settings

To train the network model, numerous satellite images were procured. (1) Google Earth images, totaling 100,000, were acquired from Google Earth Pro. Images spanning different periods were utilized as corresponding image pairs, with 80,000 designated for training, 10,000 for validation, and the remaining 10,000 for testing. (2) Multiphase images captured by WorldView-2 were employed as matching pairs, with a similar distribution of 100,000 images in total, downloaded from the geospatial data cloud. (3) SAR-optical images, also summing up to 100,000, were obtained from the SpaceNet (Van et al., 2018). These data were manually selected by choosing image pairs from the geospatial data cloud.

Fig. 5 delineates the method for generating sample data, where (a), (b) and (c) are the reference image, sensed image, and transformed image, respectively, which are aligned at a pixel level. The sensed image is distorted using a random projection transformation matrix, wherein the displacement of the x and y coordinates of the four corner points of the images is randomized between 0 and 20, to obtain the transformed


$\begin{bmatrix} x' \\ y' \\ 1 \end{bmatrix} = \begin{pmatrix} \theta_{11} & \theta_{12} & \theta_{13} \\ \theta_{21} & \theta_{22} & \theta_{23} \\ \theta_{31} & \theta_{32} & \theta_{33} \end{pmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix}$
Ground truth
projection transformation matrix

(a) Reference image (b) Sensed image (c) Warped sensed image

Fig. 5. Training dataset generation process: (a) reference image, (b) sensed image, and (c) warped sensed image.

image. In Fig. 5, (a) and (c) represent the training images, while the random projection transformation matrix is referred to as the ground truth.

Training samples tiled to 256 × 256 pixels are produced from each of the three datasets by utilizing the aforementioned approach. The threshold for coarse matching confidence, $P_c(i,j)$, is set at 0.7. All experiments in this paper are performed on an RTX3060 GPU. The training process employed the Adam optimizer with an initial learning rate of 0.0001, lasting a total of 300 epochs.

4.3. Matching performance

This section details the performance of DF-Net with respect to comparative methods on three test datasets. DF-Net is compared with traditional feature matching methods (SIFT, Oriented FAST and Rotated BRIEF (ORB) (Rublee et al., 2011), SAR-SIFT, Position Scale Orientation SIFT (PSO-SIFT) (Ma et al., 2016)) and representative deep learning local descriptor methods HardNet (Mishchuk et al., 2017), LoFTR. The average Rep., MMA, RMSE and the number of matching images at different pixel thresholds are used as evaluation metrics. Figs. 6 and 7 show the Rep., MMA at different pixel thresholds on the three test datasets, respectively. Overall, the Rep., MMA curves of DF-Net at different pixel thresholds are higher than those of the compared methods. For the Google Earth dataset, DF-Net has higher Rep. within 3 to 9 pixels thresholds and has roughly the same MMA as LoFTR. For the WorldView-2 dataset, DF-Net has an overall higher Rep., MMA curve than the compared methods. For the SAR-optical dataset, the Rep., MMA curves of SIFT and ORB are almost constant, and no match can be obtained for the SAR-optical datasets. SAR-SIFT and PSO-SIFT results are overall lower than the learning-based methods HardNet and LoFTR. Moreover, the MMA score achieved by DF-Net surpasses that of the compared methods across all pixel thresholds. At pixel thresholds $px > 5$, Rep. of DF-Net is generally superior to that of HardNet and LoFTR. When pixel threshold $px = 1$, DF-Net demonstrates higher Rep. and MMA, signifying that it has greater precision and accuracy.

Table 1 lists the total number of matches (n) and the average RMSE on the three test datasets. For the Google Earth dataset, all methods obtain close to sub-pixel RMSE. The n of ORB is larger, while the RMSE is overall higher than all compared methods. The n of HardNet is lower than LoFTR and DF-Net, and has a larger RMSE. The n of ORB is second to LoFTR, higher than SIFT and PSO-SIFT, while PSO-SIFT has a lower RMSE than other traditional feature matching methods. For the WorldView-2 and SAR-optical datasets, DF-Net gives a smaller RMSE. For the SAR-optical dataset, DF-Net's RMSE is significantly lower than the other compared methods.

Fig. 8 depicts the qualitative matching outcomes of our proposed method, demonstrating randomly chosen image pairs from the three datasets represented by G_i, W_i and S_i. The first column is Google Earth

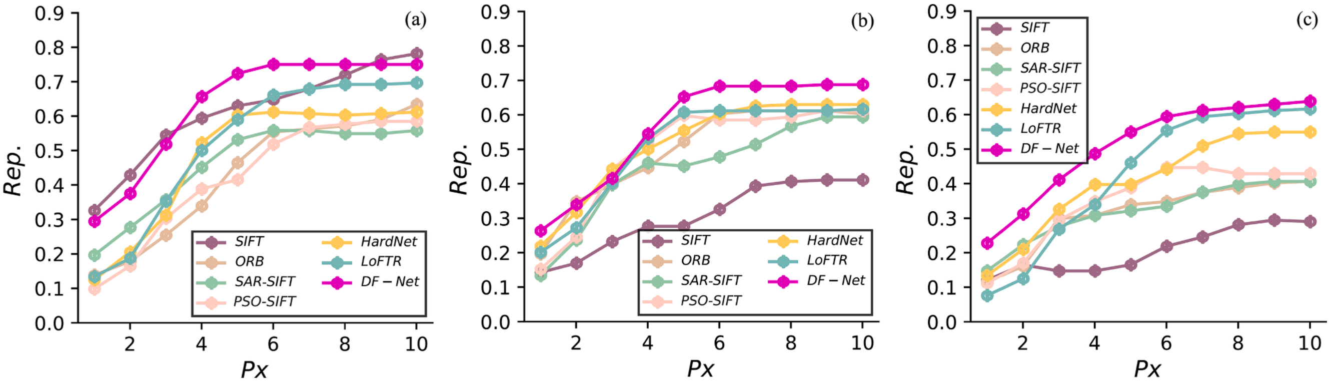


Fig. 6. Evaluation (Rep.) on three datasets: (a) Google Earth, (b)WorldView-2, and (c)SAR-optical image.

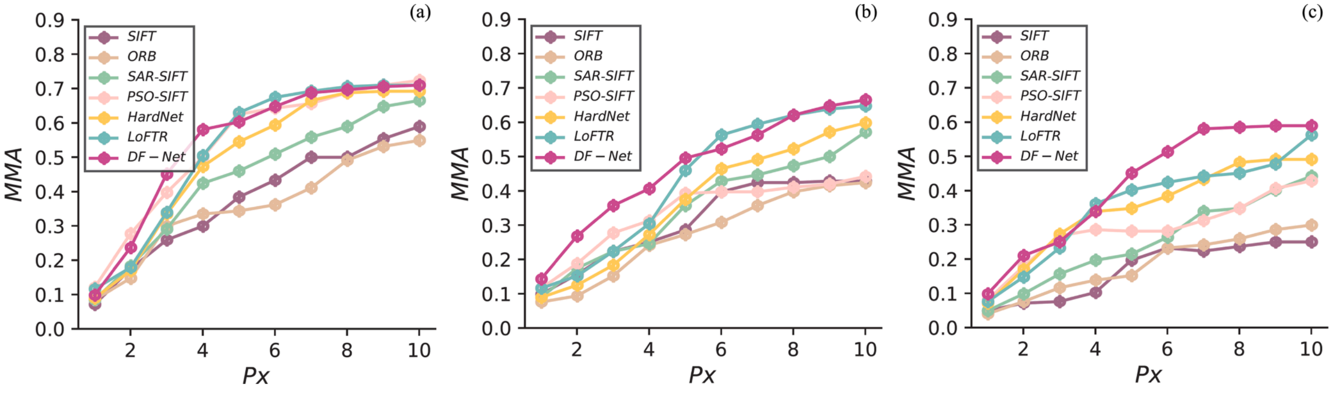


Fig. 7. Evaluation (MMA) on three datasets: (a) Google Earth, (b)WorldView-2, and (c)SAR-optical image.

Table 1
Matching results of the comparison method and DF-Net on the three datasets.

Methods	Google Earth dataset		WorldView-2 dataset		SAR-optical dataset	
	N o. of points	RMSE (*pixel*)	N o. of points	RMSE (*pixel*)	N o. of points	RMSE (*pixel*)
SIFT	550	1.49	591	6.70	120	9.93
ORB	1721	1.92	1520	5.24	1533	12.7
SAR-SIFT	670	1.53	560	4.31	430	5.79
PSO-SIFT	450	1.43	740	3.21	510	6.56
HardNet	650	1.48	603	3.25	430	4.76
LoFTR	850	1.37	640	2.53	550	4.62
DF-Net	651	1.29	610	2.32	553	3.58

images, the second column is WorldView-2 images, and the third column is SAR-optical images. It shows that for Google Earth and WorldView-2 images, the correspondences obtained by DF-Net are almost uniformly distributed over the images. In particular, for G_3, G_6, which contains more indistinguishable low-texture regions, i.e., waters, DF-Net still obtains a good correspondence over this region. This is owed to the proposed vertical self-attention module, which obtains a larger range of semantic features, making the feature descriptions obtained by the network distinguishable. For the SAR-optical image pairs, the obtained matching correspondence is overall lower than that of the Google Earth and WorldView-2 image pairs. However, the correspondence can fully register the SAR-optical images.

Fig. 9 presents a qualitative comparison of pixel alignment across three test datasets. To visualize these registration images, we superimpose them and select a distinct object (indicated by red marked boxes) that exhibits a characteristic matching result, serving as an evaluation criterion for alignment accuracy. In the case of the Google Earth image, buildings appear misaligned due to varying shooting angles; however, other features such as roads exhibit remarkable overlap. For both WorldView-2 and SAR-optical images, DF-Net achieves superior alignment in all selected areas. Although other matching methods demonstrate satisfactory alignment within the chosen areas of the Google Earth image, their performance significantly lags behind DF-Net for WorldView-2 and SAR-optical images. This comprehensive quantitative and qualitative assessment substantiates the efficacy of DF-Net for image registration.

4.4. Large-scale satellite image matching performance

In the above experiments, the effectiveness of DF-Net is evaluated. However, these tests are only limited to matching images with the size of 256 × 256 pixels. Consequently, to comprehensively assess the registration performance of DF-Net on large-scale images, we use it to establish correspondences for large-size satellite images. Corresponding to the three datasets, we select three pairs of satellite images with large scenes, denoted by I_1, I_2 and I_3, respectively. I_1 is a scene from Google Earth with two images, both of size 1478 × 1191 pixels, captured in the urban scene of Atlanta, Georgia, USA, with a resolution of 0.53 m. I_2 is a scene taken by WorldView-2 over Tripoli, Libya, with two images of size 1363 × 1053 pixels, respectively. I_3 is a large-scene SAR-optical with two images obtained by stitching from the SpaceNet dataset, with sizes of 1230 × 930 pixels. We manually collected the keypoints to determine the correspondence of these images for the evaluation of the registration.

Since DF-Net is optimized on images with the size of 256 × 256 pixels, large size satellite images cannot be directly input into the network to obtain matching. Therefore, we first superimpose the two images, crop the 256 × 256 pixels size image block at the corresponding position for obtaining matching correspondence, and slide until the correspondence of the whole image is obtained. To eliminate more errors, we use RANdom SAmple Consensus (RANSAC) to constrain outlier points. In the experiments, DF-Net is compared with SIFT, ORB, SAR-SIFT, PSO-SIFT, HardNet, LoFTR. Rep., MMA with two-pixel thresholds and RMSE is utilized to assess the performance of the matching process.

Quantitative results. Table 2 shows the results on three large size satellite images. Larger Rep., MMA and smaller RMSE indicate higher matching accuracy and precision. The results show that DF-Net can match all images. The Rep., MMA and RMSE obtained on three pairs of images, I_1 are 0.83, 0.79, 1.46, I_2 are 0.78, 0.67, 2.08 and I_3 are 0.62, 0.57, 3.49, respectively. DF-Net achieves the best values of Rep., MMA and RMSE on I_2, I_3. Additionally, it achieves comparable results on I_1. This illustrates the effectiveness of DF-Net for cross-modal matching. SIFT and ORB cannot obtain a match on I_3 at all, obtaining the worst matching results, and many experiments also demonstrate the limitations of SIFT and ORB in cross-modal matching. On I_3, PSO-SIFT employs a range of constraints, yielding lower RMSE values compared to SIFT and ORB; however, the overall performance remains unsatisfactory. SAR-SIFT is sensitive to multimodal satellite images, and the key features obtained are not reproducible. The Rep., MMA, and RMSE of

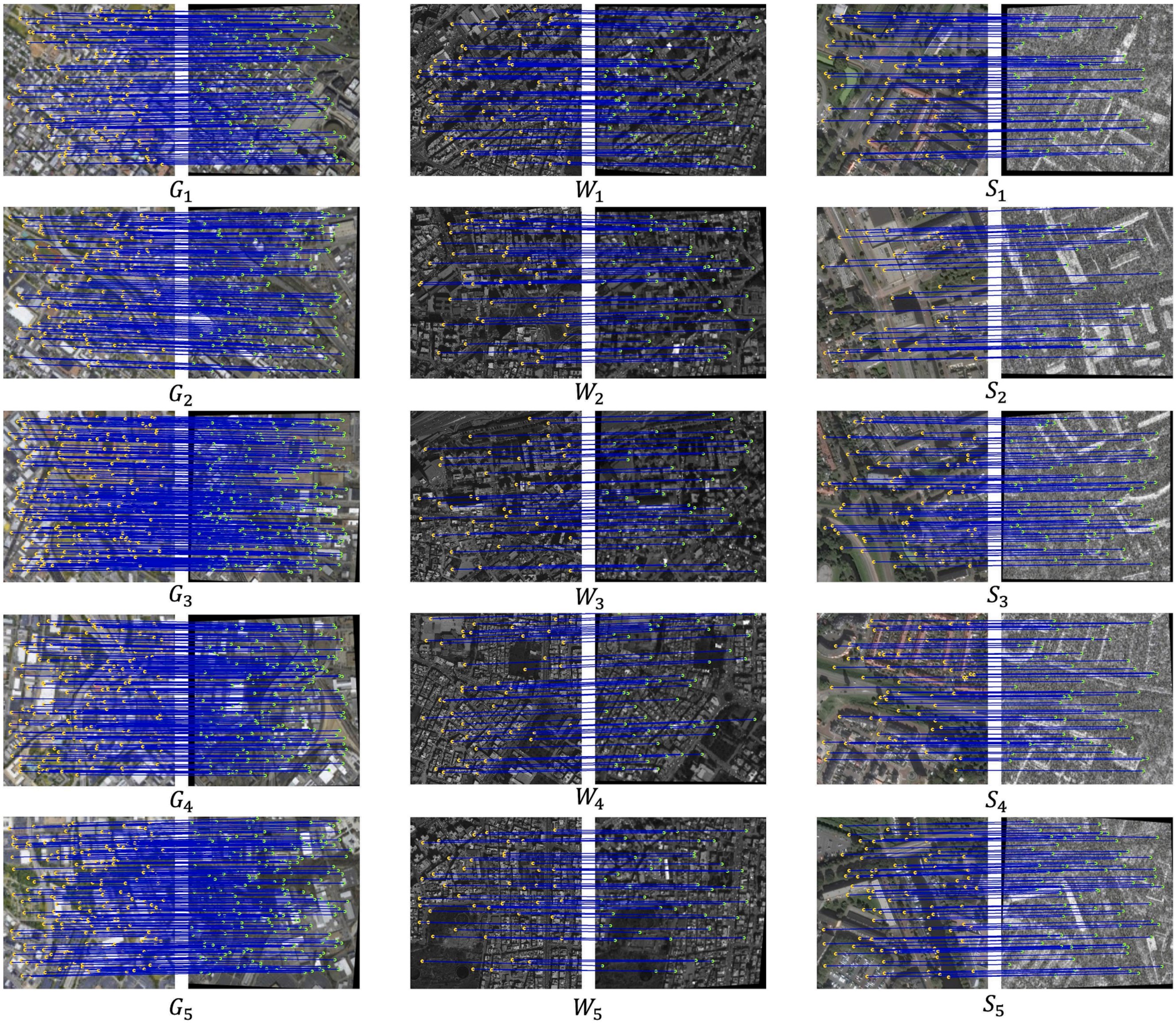


Fig. 8. Qualitative matching results of DF-Net on three datasets. These images are selected randomly, with G_i, W_i, S_i denoting different image pairs: Google Earth, WorldView-2, and SAR-optical images.

HardNet, LoFTR on I_3 are generally lower than those of DF-Net. This may be because the feature descriptions obtained by HardNet and LoFTR do not have cross-modal similarity in feature detection and description. HardNet and LoFTR obtain fewer matching correspondences and their local features are not distinguishable.

Qualitative comparisons. Fig. 10 displays the pixel superposition maps after establishing matching points and aligning each image pair. DF-Net can obtain a greater number of matching correspondences for the images across the three large-scale scenes. The superimposed maps reveal that the pixels exhibit continuity and smooth edges. The quantitative and qualitative assessments of our proposed matching method's effectiveness showcase its robust performance in large-scale satellite images.

4.5. Ablation study on feature map resolution

In our experiments, we first perform coarse matching on the low-resolution feature map, which is followed immediately by fine matching on the high-resolution feature map for each coarse match. Consequently, the size of the low-resolution feature map influences the number of matches obtained. For example, the maximum number of matches obtained on a coarse-level feature map of 1/8 is $(256/8) \times (256/8)$, while the maximum number obtained for a coarse-level feature map of 1/16 is $(256/16) \times (256/16)$. Lower resolution feature maps enable the aggregation of more distinguishable features. However, due to the small overlap between the two images, namely the size of their shared view, the obtained correspondences cannot fully establish a correspondence between the two images. During the fine matching process, increasing the resolution results in longer network training and matching prediction times. Therefore, it is necessary to choose a suitable matching resolution to ensure high matching accuracy. To evaluate the impact of resolution size on matching accuracy, we selected resolutions of 1/4, 1/8, and 1/16 for coarse matching, and 1/1 and 1/2 for fine matching. Moreover, we performed an exhaustive combination of coarse and fine matching resolutions, denoted by $c_1 = (1/4, 1/1)$, $c_2 = (1/4, 1/2)$, $c_3 = (1/8, 1/1)$, $c_4 = (1/8, 1/2)$, $c_5 = (1/16, 1/1)$ and $c_6 = (1/16, 1/2)$, on the test dataset to evaluate the of the effect of resolution on performance and speed. To ensure fairness, we keep the learning rate and batch size constant across these combinations.

Table 3 provides a comprehensive evaluation of the matching

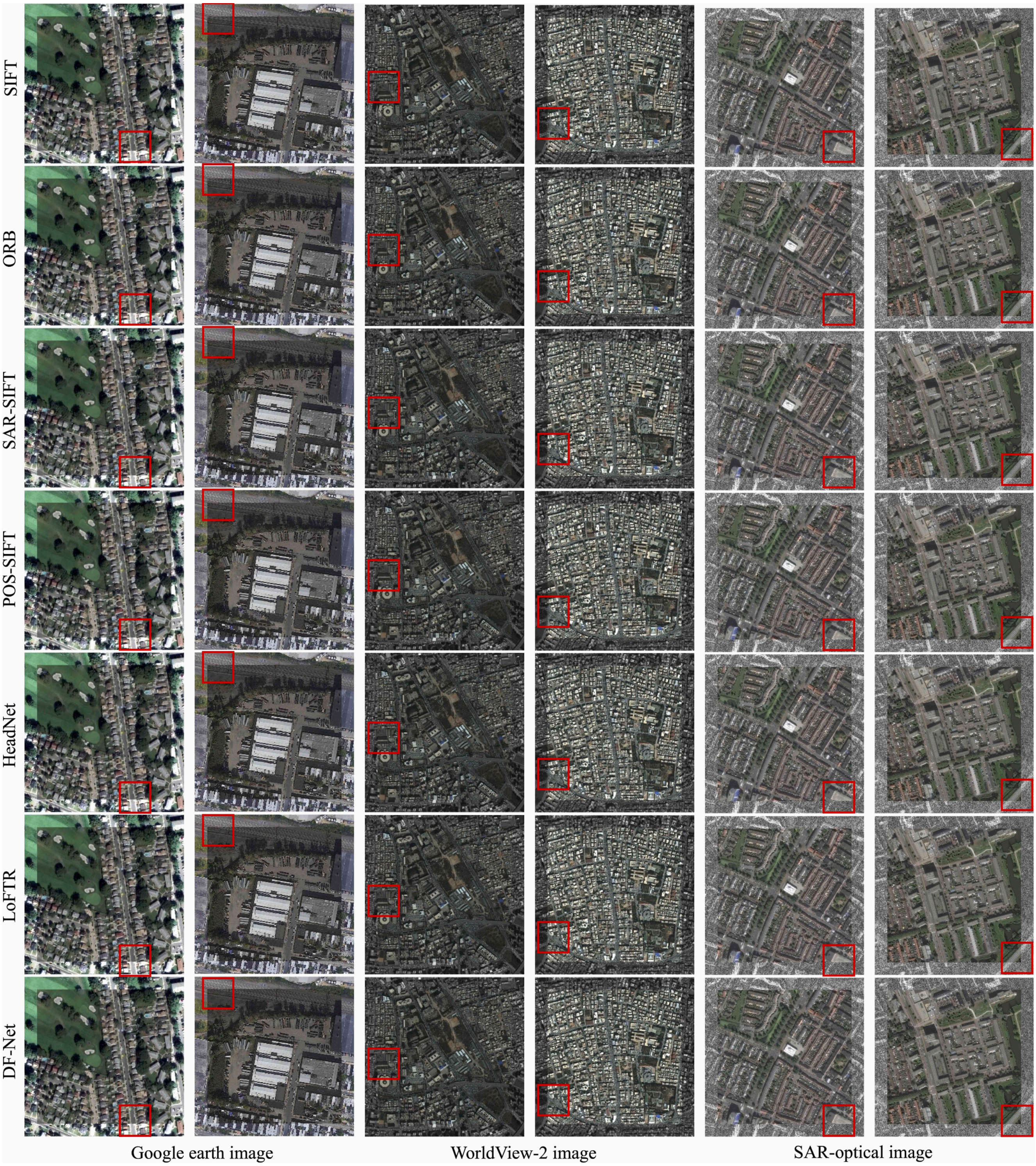


Fig. 9. Qualitative comparison of pixel alignment on three test datasets: Google Earth, WorldView-2, and SAR-optical images (red boxes: the objects of interest). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

performance across various resolution combinations, c_1, c_2, c_3, c_4, c_5 and c_6, on the Google Earth, WorldView-2, and SAR-optical test datasets. From our analysis, several critical observations emerge: Firstly, c_1, c_2, and c_3 consistently outperform other combinations in terms of Rep. and MMA metrics and exhibit the lowest RMSE across the three test datasets. This indicates that these combinations are especially effective in capturing and matching distinctive features, resulting in more accurate image correspondences. However, it's pivotal to note that even though they excel, the margin of superiority over other combinations isn't overwhelmingly vast. Secondly, c_6 standout feature is its efficiency. It consistently demonstrates the shortest mean time consumed (T), highlighting its potential for applications where processing speed is a priority, albeit at a slight accuracy trade-off. Furthermore, a dataset-specific trend was observed. For the Google Earth and WorldView-2 datasets, c_1 and c_2 took the lead in Rep., while MMA was dominated by c_3 and c_4. In contrast, the SAR-optical dataset showed a preference for

Table 2
Matching results of the comparison method and DF-Net on three large-scene satellite images.

Methods	Google Earth dataset			WorldView-2 dataset			SAR-optical dataset		
	Rep. (%)	MMA(%)	RMSE (*pixel*)	Rep. (%)	MMA(%)	RMSE (*pixel*)	Rep. (%)	MMA(%)	RMSE (*pixel*)
SIFT	85	77	1.54	50	47	5.46	46	42	9.59
ORB	74	78	1.90	66	36	6.40	31	26	10.68
SAR-SIFT	83	72	1.42	65	65	4.40	41	38	8.82
PSO-SIFT	82	77	1.81	63	59	3.93	46	41	6.52
HeadNet	81	73	1.29	51	46	2.56	53	42	4.81
LoFTR	85	77	1.97	54	50	2.10	43	36	3.95
DF-Net	83	79	1.46	68	60	2.08	62	57	3.49

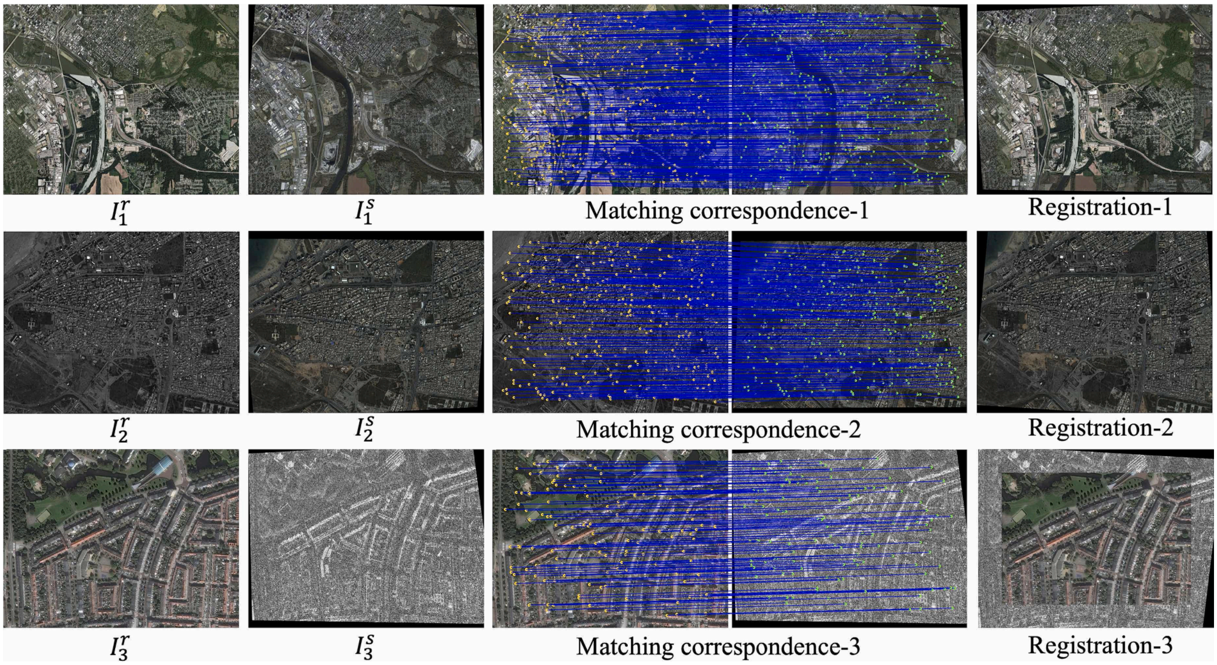


Fig. 10. Qualitative matching results on three large-size images: I_1(Google Earth) I_2 (WorldView-2) and I_3 (SAR-optical).

Table 3
Rep., MMA(2*px*) and RMSE for evaluating the matching performance on Google earth, WorldView-2 and SAR-optical datasets.

	Google Earth dataset				WorldView-2 dataset				SAR-optical dataset			
	Rep. (%)	MMA(%)	RMSE (*pixel*)	*T*(s)	Rep. (%)	MMA(%)	RMSE (*pixel*)	*T*(s)	Rep. (%)	MMA(%)	RMSE (*pixel*)	*T*(s)
c_1	79	67	1.60	13.4	76	63	2.29	13.0	60	42	3.68	13.4
c_2	82	72	1.61	9.96	74	64	2.27	9.38	67	44	3.70	9.94
c_3	78	69	1.68	2.73	67	67	2.21	5.85	67	49	3.67	5.62
c_4	71	73	1.43	3.59	57	54	2.49	3.93	66	73	3.67	3.88
c_5	70	66	1.51	3.16	59	46	2.94	3.20	68	57	3.73	3.19
c_6	72	65	1.5	1.46	61	57	2.46	1.64	59	39	3.66	1.54

c_4 in MMA and c_5 in Rep. This suggests that the optimal resolution combination may be contingent on the intrinsic characteristics of the datasets being evaluated. It's crucial to underscore that while c_1, c_2, and c_3 consistently delivered superior results, the incremental gains in performance weren't exponentially better than their counterparts. In fact, when juxtaposing performance and processing speed, we found c_4 to offer a balanced compromise, making it our choice for subsequent tests.

4.6. Ablation study on coarse matching thresholds

The choice of the confidence matrix threshold for coarse matches is crucial. Ideally, the confidence value for all possible matches in the shared view is 1.0, with possible values lying in the range [0,1]. A larger threshold value obtains less correspondence, which may not satisfy the overall matching requirements. Reducing the threshold values leads to a greater number of coarse matches, and potentially false matches, but this also increases the computational time required for fine matching. Consequently, threshold values of $t > 0.4$, $t > 0.5$, $t > 0.6$, $t > 0.7$, $t > 0.8$, and $t > 0.9$ are selected to assess the overall matching performance on the three test datasets. The number of matches (n), Rep., MMA, and RMSE serve as evaluation metrics for this analysis.

Table 4 presents the experimental outcomes at various thresholds for the three datasets. $t > 0.4$ obtains a higher number of correspondences on all three datasets, while Rep. is the lowest overall. For Google Earth, $t > 0.5$ and $t > 0.6$ perform similarly in Rep., MMA and RMSE. The highest Rep. is obtained at a threshold of $t > 0.8$ and the highest RMSE is

Table 4
The number of matches (n), Rep., MMA, and RMSE for evaluating the matching performance on different matching confidence thresholds on Google earth, WorldView-2, and SAR-optical datasets.

	Google Earth dataset				WorldView-2 dataset				SAR-optical dataset			
	N umber of point	Rep. (%)	MMA(%)	RMSE (*pixel*)	N umber of point	Rep. (%)	MMA(%)	RMSE (*pixel*)	N umber of point	Rep. (%)	MMA (%)	RMSE (*pixel*)
$t > 0.4$	737	77	70	1.55	656	66	40	2.28	495	51	60	3.81
$t > 0.5$	683	61	55	1.44	546	71	59	2.55	481	54	52	3.95
$t > 0.6$	688	78	79	1.25	491	69	65	2.44	210	61	62	3.98
$t > 0.7$	625	72	60	1.57	366	75	66	2.34	315	54	47	3.80
$t > 0.8$	572	77	73	1.51	397	70	46	2.86	305	56	63	3.72
$t > 0.9$	552	83	54	1.34	450	69	48	2.65	436	63	37	3.93

obtained at a threshold of $t > 0.9$. For WorldView-2, the highest Rep., MMA, and RMSE are obtained at a threshold of $t > 0.8$, however, n is smaller. For the SAR-optical dataset, a threshold of $t > 0.9$ obtained the highest Rep., MMA and RMSE values at $t > 0.8$. Under different threshold conditions, overall, a threshold value of $t > 0.4$ obtained a higher number with smaller Rep., MMA, and RMSE. In contrast, at a threshold value of $t > 0.9$, higher Rep., MMA, and RMSE are obtained on the three datasets, while n is smaller.

Higher values of n, Rep., and MMA indicate better matching performance, while a smaller RMSE denotes greater matching accuracy. To ensure that larger values for all metrics represent increased accuracy, we utilized the inverse of RMSE. The entropy weighting method is employed to assign weights to each indicator, obtaining indicator score values for all thresholds across the three datasets. As shown in Fig. 11, for the Google Earth and WorldView-2 datasets, the highest score is achieved with a threshold value of $t > 0.8$. For the SAR-optical dataset, the optimal threshold is $t > 0.7$, yielding the highest metric score value. Consequently, in subsequent experiments, we assessed the test datasets using different thresholds for evaluation, respectively.

4.7. Ablation study on network configuration

To facilitate the design of the proposed model, we perform ablation studies to investigate the influence of integrating different network architectures on the matching accuracy. Critical operations for achieving coarse-to-fine matching of the two images involve coarse and fine matching processes. Therefore, we employ convolutions with a substantial number of parameters (for obtaining 1/2 and 1/8 size feature maps) and coarse-to-fine matching operations as the base network (CNNBase). We experiment with a combination of parallel ResNet, vertical self-attention, and the number of self-attention layers. All networks are trained and tested using three datasets. We maintain determinism in the random data during training. The matching performance is evaluated using Rep., MMA, and RMSE as metrics. The results on the three datasets are presented in Table 5.

Table 5 depicts the matching results obtained on the three datasets, and as expected, the best matching performance is obtained using parallel ResNet and vertical self-attention networks. When using only *CNNBase*, the average Rep. on the three datasets is 0.57, MMA is 0.52, and RMSE is 3.34, respectively, with the lowest overall evaluation metric values. The matching performance improves by replacing the convolution in the *CNNBase* with ResNet, with the average Rep., MMA improving significantly to 0.66, 0.59 and the average RMSE decreasing to 3.43, respectively. When the ResNet and self-attention are combined to test DF-Net, the average Rep., MMA improves from 0.66 ± 0.04, 0.59 to 0.75, 0.53 on the three datasets, respectively. The average RMSE decreases from 3.43 to 2.38, especially apparent on the SAR-optical dataset. These experiments demonstrate the effectiveness of the proposed vertical self-attention module for improving matching performance. This module increases the larger receptive field in the feature extraction process, while facilitating the fusion of information between two images for obtaining feature descriptions with similarity. In addition, we increase the number of layers for vertical self-attention to $n = 6$, 8. On the three datasets, the average Rep., MMA increase insignificantly, and similarly the average RMSE decrease by 0.11. Therefore, we select $n = 4$ as the number of vertical self-attention layers. The subsequent experiments are based on the parallel ResNet, vertical self-attention module. Fig. 12 depicts the RMSE of the combined network on the validation dataset. Overall, the RMSE curve for ResNet + self-attention ($layer = 4$) is lower than the other networks during the training process, which is consistent with the results obtained in Table 5.

4.8. Ablation study on coarse to fine-grained mapping

To provide a deeper insight into the impact of the transition from coarse to fine-grained mapping, we performed an additional ablation experiment. The focus of this study was to understand the computational cost introduced by the fine-grained mapping and to assess the potential performance gains in the matching accuracy. Additionally, the DF-Net's complexity can be gauged using two measures: model parameter size

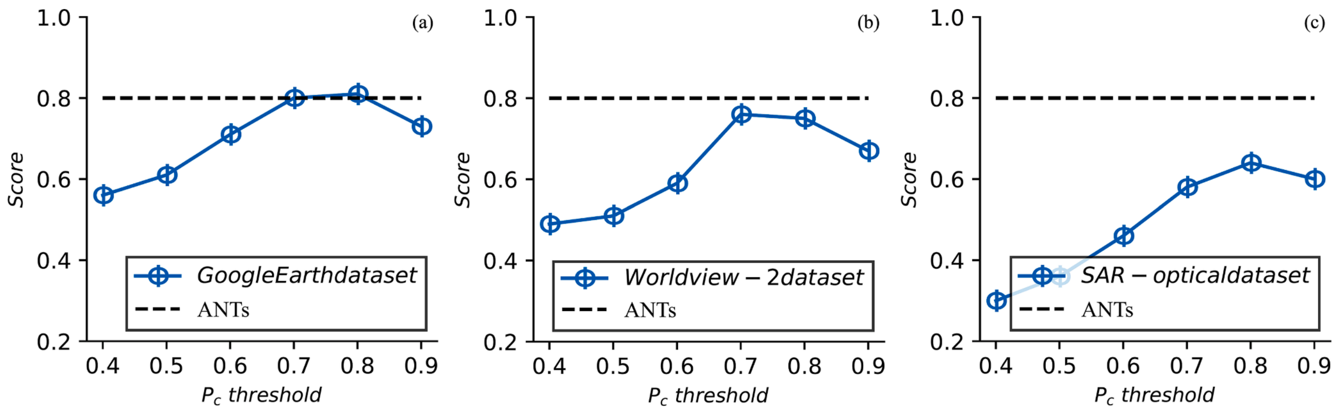


Fig. 11. Score results under different matching confidence thresholds: (a) Google Earth image (b)Worldview-2 image, and (c)SAR-optical image.

Table 5
Ablation study results. RSA, RSA (l = 6)and RSA (l = 8) correspond to ResNet + Self-attention, ResNet + Self-attention (layer = 6) and ResNet + Self-attention (layer = 8).

	Google Earth dataset			WorldView-2 dataset			SAR-optical dataset		
	Rep. (%)	MMA(%)	RMSE (*pixel*)	Rep. (%)	MMA(%)	RMSE (*pixel*)	Rep. (%)	MMA(%)	RMSE (*pixel*)
CNNBase	64	58	2.17	62	52	3.73	46	47	4.42
ResNet	75	63	2.07	68	59	3.82	58	55	4.39
RSA	84	73	1.26	77	64	2.39	67	58	3.51
RSA (l=6)	82	73	1.52	77	64	2.40	67	58	3.50
RSA (l = 8)	82	72	1.44	76	62	2.41	63	58	3.52

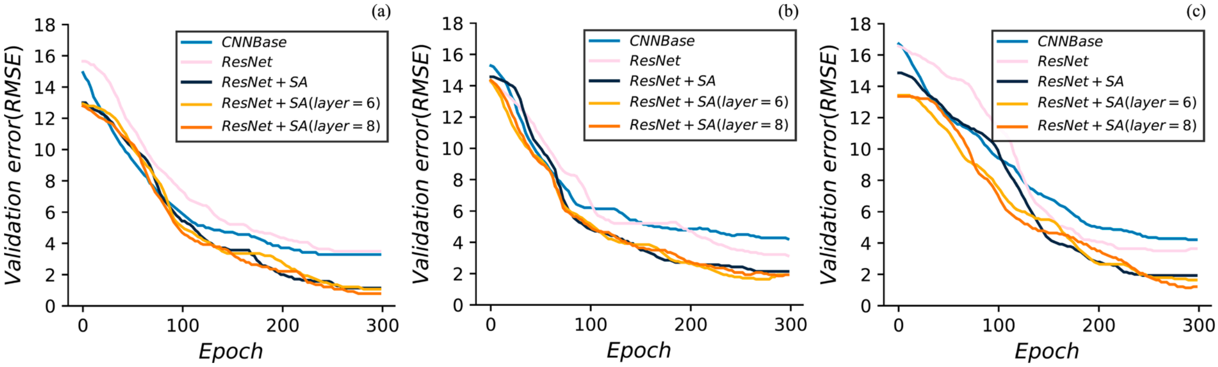


Fig. 12. RMSE of the combined network for the ablation study on three validation datasets: (a) Google Earth, (b)Worldview-2, and (c)SAR-optical dataset.

and the Floating Point Operations Per Second (FLOPs). Analysis reveals that the DF-Net possesses a model size of 198 MB and demands a computational capacity of 6 GFLOPs.

In this ablation study, we evaluated two scenarios: (1) utilizing only coarse mapping without transitioning to fine-grained mapping and (2) the complete process involving a transition from coarse to fine-grained mapping. This would allow us to identify the additional computational overhead introduced by the fine-grained mapping and the subsequent improvement in performance metrics.

As shown in Table 6, the transition from coarse to fine-grained mapping introduces a computational overhead, increasing the computation time from 7.5 s to 13.2 s. However, this transition also leads to a significant improvement in the matching accuracy metrics, with the Rep. increasing from 65 % to 82 % and the MMA from 56 % to 73 %. The RMSE also saw a considerable reduction, indicating improved matching precision. Therefore, the transition from coarse to fine-grained mapping, though computationally more intensive, provides significant performance gains, making it an essential step in the proposed approach.

5. Conclusion

In this paper, we have proposed a dual-branch fusion network (DF-Net) to rectify the limitations posed by the non-repeatability of keypoints procured from contemporary cross-modal detectors. This network was predicated on a coarse-to-fine matching strategy designed to facilitate the establishment of correspondences between satellite imagery, avoiding the conventional sequential steps of detection, description, matching, and geometric constraint. Both reference and sensed images were deployed as inputs for the dual-branch matching network, subsequently producing feature descriptions of both high and low resolution for the process of coarse-to-fine matching. A cross-attention mechanism was introduced to interactively combine the outputs from each stage in the dual-branch networks. The objective was to obtain feature descriptions that included each other's data, culminating in the acquisition of cross-modal feature similarity descriptions. Experimental trials affirmed the efficacy of the proposed method, evidencing exceptional matching accuracy. It is worth noting that the average Rep., MMA, and RMSE of the DF-Net applied to three large-scale satellite images registered at 0.71, 0.65, and 2.34, respectively. This underscores the DF-Net's effectiveness with satellite images, which yielded a performance that was highly competitive.

Table 6
Ablation study results on the transition from coarse to fine-grained mapping.

Scenario	Rep. (%)	MMA (%)	RMSE (pixel)	Computation Time (s)
Coarse Mapping Only	65	56	2.5	7.5
Coarse to Fine-Grained Mapping	82	73	1.3	13.2

CRediT authorship contribution statement

Liangzhi Li: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing – original draft, Writing – review & editing, Visualization. **Ling Han:** Investigation, Visualization, Data curation. **Kyle Gao:** . **Hongjie He:** Investigation, Data curation, Funding acquisition. **Lanying Wang:** Investigation. **Jonathan Li:** Resources, Writing – review & editing, Supervision, Project administration.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The authors do not have permission to share data.

Acknowledgments

This work was supported in part by the National Natural Science

Foundation of China under Grant 42171348 and a doctoral scholarship provided by the China Scholarship Council (CSC).

References

Cangea, C., Veličković, P., Lio, P., 2019. Xflow: Cross-modal deep neural networks for audiovisual classification. IEEE Trans. Neur. Netw. Lear. Syst. 31 (9), 3711–3720.

Chen, J., Cheng, B., Zhang, X., 2022b. A TIR-visible automatic registration and geometric correction method for SDGSAT-1 thermal infrared image based on modified RIFT. Remote Sens. 14 (6), 1393.

Chen, J., Jie, T., Noah, L., Jian, Z., R Theodore, S., and Andrew F, L., 2010. A partial intensity invariant feature descriptor for multimodal retinal image registration. *IEEE Trans. Biomed. Eng.* 57 (7): 1707–1718.

Chen, H., Luo, Z., Zhou, L., Tian, Y., Zhen, M., Fang, T., ... & Quan, L. (2022). Aspanformer: Detector-free image matching with adaptive span transformer. In: *Proc. ECCV*, pp. 20-36.

Cui, Z., Qi, W., Liu, Y., 2020. A fast image template matching algorithm based on normalized cross correlation. J. Phys.: Conf. Ser. 1693 (1), 012163.

Deng, X., Liu, E., Li, S., 2023. Interpretable multi-modal image registration network based on disentangled convolutional sparse coding. IEEE Trans. Image Process. 32, 1078–1091.

Fu, Y., Yang, L., Tonghe, W., Walter J, C., Tian L., and Xiaofeng Y., 2020. Deep learning in medical image registration: A review. *Phys. Med. Biol.* 65 (20): 20TR01.

Gao, C., Li, W., Tao, R., 2022. MS-HLMO: Multiscale histogram of local main orientation for remote sensing image registration. IEEE Trans. Geosci. Remote Sens. https://doi.org/10.1109/TGRS.2022.3193109.

Geigle, G., Pfeiffer, J., Reimers, N., Vulić, I., Gurevych, I., 2022. Retrieve fast, rerank smart: Cooperative and joint approaches for improved cross-modal retrieval. Trans. Asso. Compu. Lingu. 10, 503–521.

Haskins, G., Kruger, U., Yan, P., 2020. Deep learning in medical image registration: A survey. Machine Vis. Appl. 31, 1–18.

He, K., Zhang, X., Ren, S., and Sun, J., 2016. Deep residual learning for image recognition. In: *Proc. CVPR*, pp. 770–778.

He, Y., Xiang, S., Kang, C., Wang, J., Pan, C., 2016b. Cross-modal retrieval via deep and bidirectional representation learning. IEEE Trans Multim. 18 (7), 1363–1377.

Jiang, X., Ma, J., Jiang, J., 2019. Robust feature matching using spatial clustering with heavy outliers. IEEE Trans. Image Process. 29, 736–746.

Jiang, B., Sun, P., Luo, B., 2022. GLMNet: Graph learning-matching convolutional networks for feature matching. Patt. Recog. 121, 108167.

Khowaja, S.A., Lee, S.L., 2020. Hybrid and hierarchical fusion networks: a deep cross-modal learning architecture for action recognition. Neur. Compu. Appl. 32, 10423–10434.

Li, L., Liu, M., Ma, L., 2022. Cross-Modal feature description for remote sensing image matching. Int. J. Appl. Earth Obs. Geoinf. 112, 102964.

Li, L., Han, L., Ding, M., Cao, H., 2023. Multimodal image fusion framework for end-to-end remote sensing image registration. IEEE Trans. Geosci. Remote Sens. https://doi.org/10.1109/TGRS.2023.3247642.

Liu, Y., Xia, C., Zhu, X., 2021. Two-stage copy-move forgery detection with self deep matching and proposal superglue. IEEE Trans. Image Process. 31, 541–555.

Liu, J., Yang, M., Li, C., Xu, R., 2020. Improving cross-modal image-text retrieval with teacher-student learning. *IEEE Trans.* Circ. *Syst. Video Tech.* 31 (8), 3242–3253.

Liu, M., Zhou, G., Ma, L., 2023. SIFNet: A self-attention interaction fusion network for multisource satellite imagery template matching. Int. J. Appl. Earth Obs. Geoinf. 118, 103247.

Long, J., Evan, S., and Trevor, D., 2015. Fully convolutional networks for semantic segmentation. In: *Proc. CVPR*, pp. 3431–3440.

Lu, X., Yan, Y., Kang, B., and Du, S. 2023. ParaFormer: Parallel attention transformer for efficient feature matching. arXiv preprint arXiv:2303.00941.

Ma, J., Li, Z., Zhang, K., 2022. Robust feature matching via neighborhood manifold representation consensus. ISPRS J. Photogram. Remote Sens. 183, 196–209.

Ma, W., Wen, Z., Wu, Y., Jiao, L., Gong, M., Zheng, Y., Liu, L., 2016. Remote sensing image registration with modified SIFT and enhanced feature matching. IEEE Geosci Remote Sens. Lett. 14, 3–7.

Meng, L., Zhou, J., Liu, S., 2021. Investigation and evaluation of algorithms for unmanned aerial vehicle multispectral image registration. Int. J. Appl. Earth Obs. Geoinf. 102, 102403.

Mishchuk, A., Dmytro, M., Filip, R., and Jiri, M., 2017. Working hard to know your neighbor's margins: Local descriptor learning loss. In: *Proc. NeurIPS,* pp. 4828-4840.

Ng, P.C., Steven, H., 2003. SIFT: Predicting amino acid changes that affect protein function. Nucleic Acids Res. 31 (13), 3812–3814.

Prakash, A., Chitta, K., and Geiger, A. 2021. Multi-modal fusion transformer for end-to-end autonomous driving. In: *Proc. CVPR*, pp. 7077-7087.

Quan, D., Wang, S., Gu, Y., 2022. Deep feature correlation learning for multi-modal remote sensing image registration. IEEE Trans. Geosci. Remote Sens. https://doi.org/10.1109/TGRS.2022.3187015.

Revaud, J., Philippe, W., César D., Noe, P., Gabriela, C., Yohann, C., and Martin, H., 2019. R2D2: repeatable and reliable detector and descriptor, arXiv preprint arXiv:1906.06195.

Rublee, E., Vincent, R., Kurt, K., and Gary, B., 2011. ORB: An efficient alternative to SIFT or SURF. In: *Proc. ICCV*, pp. 2564–2571.

Sedaghat, A., Mohammadi, N., 2019. Illumination-robust remote sensing image matching based on oriented self-similarity. ISPRS J. Photogram. Remote Sens. 153, 21–35.

Sengupta, D., Gupta, P., Biswas, A., 2022. A survey on mutual information based medical image registration algorithms. Neurocomp. 486, 174–188.

Sun, J., Shen, Z., Wang, Y., Bao, H., Zhou, X., 2021. LoFTR: Detector-free local feature matching with transformers. *Proc. CVPR* 8922–8931.

Tian, X., Shao, J., Ouyang, D., Zhu, A., and Chen, F. 2022. SMDT: Cross-view geo-localization with image alignment and transformer. In: *Proc. ICME,* pp. 1-6.

Van, Etten A., Lindenbaum, D., Bacastow, T M.,2018. Spacenet: A remote sensing dataset and challenge series. arXiv preprint arXiv:1807.01232.

Wang Q, Zhang J, Yang K, et al. Matchformer: Interleaving attention in transformers for feature matching. In: *Proc. CVPR*, 2022: 2746-2762.

Wang, M., Zhang, J., Deng, K., 2021. Combining optimized SAR-SIFT features and RD model for multisource SAR image registration. IEEE Trans. Geosci. Remote Sens. https://doi.org/10.1109/TGRS.2021.3074630.

Wang, S., Zhuang, F., Jiang, S., Huang, Q., Tian, Q., 2015. Cluster-sensitive structured correlation analysis for web cross-modal retrieval. Neurocomp. 168, 747–760.

Wei, Y., Zhao, Y., Lu, C., Wei, S., Liu, L., Zhu, Z., Yan, S., 2016. Cross-modal retrieval with CNN visual features: A n-ew baseline. IEEE Trans. Cybe. 47 (2), 449–460.

Wei, X., Zhou, L., 2021. AI-enabled cross-modal communications. *IEEE Wire. Commu.* 28 (4), 182–189.

Wong, A., and David, A., Clausi, 2007. ARRSI: Automatic registration of remote-sensing images. *IEEE Trans. Geosci. Remote Sens.* 45 (5): 1483–1493.

Wu, W., Shao, Z., Huang, X., 2022. Quantifying the sensitivity of SAR and optical images three-level fusions in land cover classification to registration errors. Int. J. Appl. Earth Obs. Geoinf. 112, 102868.

Xie, H., Zhang, Y., Qiu, J., Zhai, X., Liu, X., Yang, Y., Zhong, J., 2023. Semantics lead all: Towards unified image registration and fusion from a semantic perspective. Info. Fusion 98, 101835.

Xu, X., Lin, K., Gao, L., Lu, H., Shen, H.T., Li, X., 2020. Learning cross-modal common representations by private–shared subspaces separation. IEEE Trans. Cybe. 52 (5), 3261–3275.

Yang, Z., Tingting, D., Yang, Y., 2018. Multi-temporal remote sensing image registration using deep convolutional features. IEEE Access. 6, 38544–38555.

Yao, Y., Zhang, Y., Wan, Y., 2022. Multi-modal remote sensing image matching considering co-occurrence filter. IEEE Trans. Image Process. 31, 2584–2597.

Ye, Y., Lorenzo, B., Jie, S., Francesca, B., Qing, Z., 2019. Fast and robust matching for multimodal remote sensing image registration. IEEE Trans. Geosci. Remote Sens. 57 (11), 9059–9070.

Ye, F., Su, Y., Xiao, H., Zhao, X., Min, W., 2018. Remote sensing image registration using convolutional neural network features. IEEE Geosci. Remote Sens. Lett. 15 (2), 232–236.

Zhang, H., Lei, L., Ni, W., 2020. Optical and SAR image matching using pixelwise deep dense features. IEEE Geosci. Remote Sens. Lett. https://doi.org/10.1109/LGRS.2020.3039473.

Zhang, Y., Yao, Y., Wan, Y., 2023. Histogram of the orientation of the weighted phase descriptor for multi-modal remote sensing image matching. ISPRS J. Photogram. Remote Sens. 196, 1–15.