Engineered Science
DOI: https://dx.doi.org/10.30919/es1585

Local and Hierarchical Feature Fusion for Registration of Satellite Images Using Residual Deep Neural Network Modified with Attention Mechanism and Depth-Wise Separable Convolution

P. S. Tondewad[1, 2,*] and M. P. Dale[2]

Abstract

An accurate multi-sensor satellite image registration is a fundamental prerequisite for an optimal preprocessing step in a wide range of satellite image applications. This study proposes an innovative approach that integrates a residual deep neural network (RNN) model modified with an attention mechanism and depth-wise separable convolution (AM-DSC) layer, for hierarchical feature investigation and scale invariance feature transform (SIFT) for local features. These extracted features undergo dimensionality reduction through principal component analysis (PCA), followed by feature selection using the L2 Norm distance metric. Subsequently, random sample consequence (RANSAC) is adopted for image resampling and transformation. Considering the non-specific surface area captured, the algorithm has been evaluated with ten different datasets of synthetic aperture radar (SAR) and multispectral image pairs to demonstrate the accuracy of the proposed algorithm. This investigation was conducted using evaluation measures, including root mean square error (RMSE), RMSE cross-validation based on leave-one-out ($RMSE_{LOO}$) method, correctly matching ratio (CMR), peak signal-to-noise ratio (PSNR) and spectral similarity index measure (SSIM). The proposed methodology demonstrated superior performance, achieving a CMR of more than 40% and enhancing registration accuracy as measured by RMSE, PSNR, and SSIM.

Keywords: Satellite image registration; SAR image; Multispectral image; Scale invariance feature transform; Deep neural network; Feature fusion.
Received: 17 March 2025; Revised: 11 May 2025; Accepted: 03 June 2025.
Article type: Research article.

1. Introduction

As per the UNITED NATIONS Office for Outer Space Affairs, approximately 7,500 remote sensing satellites are available in space. The multisource and multitemporal images obtained by different satellites can contribute to diverse remote sensing (RS) image applications.[1,2] There are various types of sensors and image-capturing mechanisms, so each image provides different types of information.[3] The SAR images provide good textural and spatial information.[4] One of the key obstacles in the processing of multisource images is their accurate alignment. This alignment process is known as image registration. This process is about transforming two or more datasets into a common coordinate system.[5,6] The image registration process can be broadly classified into two types based on simultaneity: 1) Synchronous and 2) Asynchronous registration. Synchronous registration refers to the stitching and matching of concurrently captured images, which includes panoramic photography, whereas asynchronous image registration processes images captured at different time intervals by different sensors with different view angles. Asynchronous registration is incorporated into medical images, multisource, and multiresolution remote sensing image processing applications. Again, asynchronous registration can be of the same sensor or multi-sensor images.[1] Generally, images from the same sensor type display relatively consistent characteristics, while those from multiple sensors often exhibit greater diversity, increasing the complexity of aligning and registering these images. Image registration is an

[1] *Electronics and Telecommunication Engineering, AISSMS Institute of Information Technology, Savitribai Phule Pune University, Pune, Maharashtra, 411001, India*
[2] *Electronics and Telecommunication Engineering, Modern Education Society's Wadia College of Engineering, Savitribai Phule Pune University, Maharashtra, Pune, 411001, India*
**Email:* priyanka.tondewad@mescoepune.org (P. S. Tondewad)

essential step in RS image applications. This critical process is integral to a wide array of fields, encompassing disaster management, change detection, surveillance activities, image-aided navigation systems, ground target identification, weather forecasting research, and various other domains. [7-10] Over the last decade, significant research efforts have been directed towards the study of satellite image registration. Contemporary image registration techniques include both intensity-based, referred to as area-based, and feature-based RS image registration methods.[11] Intensity-based registration approaches typically transform the input images into standard grayscale intensity and compare their mutual or cross-correlation relation to accomplish image matching.[7,12] However, these methods are often prone to decreased accuracy due to the influence of factors such as image capture timing, angles, and other complex parameters on pixel intensity. This susceptibility makes it difficult to select an appropriate dataset for subsequent operations, potentially leading to impractical approaches.[3]

The second type of registration is feature-based methods. These methods usually perform the detection of common elements between two images, such as lines, points, line intersections, contours, and regional distinctive features, for accurate matching.[10,12,13] The crucial part is to find sufficient features from both images that are distinctive and resilient to noise. The most popularly used traditional algorithms are scale invariance feature transform (SIFT), oriented FAST and rotated BRIEF (ORB), speeded up robust features (SURF).[2,14,15] These methods give salient point descriptors, which are easy to implement and register most of the images with fewer variations.[10] Few researchers have noticed that SIFT descriptors are not always suitable to attain desirable results.[16] Therefore, further modifications in SIFT have focused on advanced feature selection and distinctive feature finding. Dellinger *et. al*.[17] have implemented SAR-SIFT, which redefined a gradient to obtain robust orientation and a magnitude without any impact of speckle noise. In general, SIFT algorithms are able to maintain brightness, scale invariance, and rotation, but fail to be robust to image noise and texture changes.[16,18,19,20] Also, SURF exhibits diminished efficacy when processing highly textured images and demonstrates susceptibility to alterations in viewpoint.[20,21]

Several researchers have attempted to combine area and feature-based methods, such as intensity information combined with SIFT, Mutual information (MI) combined with SIFT, and various other approaches.[16,18,20,21] Another hybrid implementation was presented for dental panoramic X-ray image registration in which wavelet-based decomposition was performed on reference and sensed images. The implementation combined intensity-based (MI) and feature-based (SIFT) methods, along with outlier removal using random sample consensus (RANSAC). In comparison with other methods, the proposed implementation resulted in a 0.785 normalized cross-correlation (NCCC) and 0.1040% percentage relative root mean square error (PRRMSE).[22] Although hybrid implementations obtained improved performance compared to SIFT alone, they are limited to capturing low-level local features.[23,24] The extraction of high-resolution complex textural features remains challenging even with advanced feature-based algorithms; consequently, convolutional neural network (CNN) based methods have emerged as a viable alternative.[20,25]

In recent years, there has been significant proficiency and success in the area of computer vision algorithms through CNN models. Given the ability of CNNs to process input data and extract more complex features than conventional methods, numerous researchers have focused intensively on deep learning methods for classification-based applications.

Furthermore, researchers started designing modified networks to achieve additional benefits. Hu *et. al*.[26] proposed an unsupervised CNN model for lung computed tomography (CT) image registration, while Chang *et al*.[27] developed a dense feature network integrated with an advanced regression model referred to GcrDfNet. In addition to this, a dense convolutional network (DenseNet) was utilized, incorporating partial transfer learning and selective parameter fine-tuning for enhanced performance. In parallel research, a Siamese network was utilized for registration of SAR and optical images, yielding an average RMSE of 3.37.[28] Additionally, Ye *et. al*.[29] introduced a novel hybrid matching technique, known as attention enhanced structural features (AESF), which employed AESF and improved the accuracy by surpassing handcrafted descriptor methods by approximately 6.5%.

Numerous experiments utilizing NNs for RS image processing have been conducted in recent years. However, NN training typically requires a substantial dataset. However, access to a large dataset is not always feasible. To address the challenge of limited data, Wang *et. al*.[30] proposed a network that learns from input images and transformed copies of them without the need for any external datasets, observing superior results compared to the traditional methods. Quan *et al*.[31] achieved improved results by training the NN using SAR images. Li *et al*.[6] fused the traditional algorithm with a deep learning architecture. The algorithm adopted a deep learning framework to identify hierarchical feature maps and then merge the FAST descriptors to accomplish image registration. Yet, it has managed to achieve RMSE 3.4180, which necessitates further improvement. Afterwards, the attention mechanism has been utilized in the field of RS image processing.[32-34] In alignment with the intended purpose of the network, one should select a network as demonstrated by Nurseitov *et al*.[35]. Li *et. al*.[36] introduced a fused max-average pooling (FMAPooling) operation alongside an enhanced channel attention mechanism (FMAttn), leveraging the two pooling functions to improve feature representation in deep neural networks (DNNs). These methods aim to augment multi-level features extracted through max pooling and average pooling, respectively. The efficacy of these proposals is validated using VGG, ResNet, and MobileNetV2 architectures on CIFAR10/100 and ImageNet100 datasets.

Experimental results indicate that FMAPooling yields an accuracy improvement of up to 1.63% compared to the baseline model, while FMAttn achieves an accuracy enhancement of up to 2.21% relative to the previous channel attention mechanism.

Chen *et. al*.[32] worked on the limitation of achieving high accuracy in multi-view RS image registration using an attention mechanism. They developed a dual attention mechanism with an RNN. This work highlighted that incorporating the proposed method enhances the network's proficiency in identifying and localizing image features. Han *et. al*.[37] presented a structure-consistent generative adversarial network to improve the structural similarity of generated outputs. This enhancement is achieved by enforcing constraints on the consistency of information within the local neighborhood, thereby preserving structural coherence more effectively. Liu and Zhang investigated the application of reinforcement learning to image registration, where the transformation matrix parameters are continuously optimized.[38] However, the approach has certain limitations, particularly in its applicability to multimodal image registration in real-world scenarios. The experimentations are conducted on a simulated registration dataset, and the effectiveness of the method on actual remote sensing images remains unexamined. The latest advancements in image registration have led to the development of innovative descriptors. For example, Wang *et al*.[39] introduced a multi-scale duty cycle descriptor and a multi-scale and multi-angle partitioned context weighted descriptor, achieving an RMSE of 3.35 in registering infrared-visible images. In another study,[40] the author developed a multi-feature alignment and matching network for SAR and optical image registration. This approach utilized momentum encoders to explore essential feature relationships, thereby adaptively adjusting handcrafted labels and enhancing network optimization robustness, while addressing the challenge of outlier removal in initial matching is still a thrust area. The cyclic registration network described by Wang *et al*.[41] and Han *et al*.[42] demonstrated successful image registration using the consistency generative adversarial network, albeit with significant computational complexity. Furthermore, Xiong *et al*.[43] employed cosine similarity template matching to effectively quantify similarities between SAR and optical images. Although this method successfully registered images, accurately predicting repetitive textures remains a challenge.

Despite decades of evolution in RS image registration, challenges persist in registering two images of dissimilar resolution.[20,44-46] To mitigate these issues, the present study proposed a local and global feature fusion approach. The proposed study incorporates both traditional and advanced feature findings through SIFT and modified RNN. Most salient features are extracted from the fused feature set using principal component analysis (PCA). Subsequently, a brute-force matcher is employed to identify the correct matches. Lastly, RANSAC is applied to reconstruct the overlapping image. The core contributions of this design are as follows: (1) Feature fusion of local SIFT and hierarchical RNN features, (2) Feature correlation enhanced by inclusion of attention mechanism in RNN, and (3) Tailored depth-wise separable convolutional layer.

2. Proposed algorithm framework

Image registration involves aligning multiple images from various sensors, time periods, and resolutions into a unified coordinate system by superimposing them on one another. There are four major steps: (1) Feature extraction, (2) Feature selection, (3) Feature matching, (4) Image transformation and reconstruction. Feature extraction is an imperative step. Here, the proposed algorithm extracts features from a pretrained RNN with an attention mechanism and depth-wise separable convolution (AM-DSC) model and the SIFT algorithm, followed by PCA-based feature selection and RANSAC for image transformation.

Fig. 1 illustrates the main steps of this algorithm, while Fig. 2 provides a detailed explanation of the identity block. The process begins by resizing the input image to 224×224 to reduce the computational complexity. These images are then normalized to suit the network model's input requirements. Next, feature extraction is performed using both the traditional SIFT method and RNN with AM-DSC. The extensive set of features is then processed through PCA for optimal feature selection. The chosen features undergo further processing with a brute force (BF) matcher. Finally, RANSAC is employed to identify accurate inliers and calculate the homographic matrix **H** for image registration. A comprehensive elucidation of feature extraction through image registration is provided as follows.

2.1 SIFT features

In SIFT, the given input image is the first scale space separated into octaves. Each generated octave is half the size of the previous one. Every octave has a progressively blurred image by Gaussian Blur operator in Eqs. (1) and (2).

$$G(e,f,\lambda) = S(e,f,\lambda) * J(e,f) \tag{1}$$

where $J(e, f)$ is input image.

$$S(e,f,\lambda) = \frac{1}{2\pi\lambda^2} e^{\frac{-(e^2+f^2)}{2\lambda^2}} \tag{2}$$

where $S(e, f, \lambda)$ is Gaussian function with scale λ.

Further Difference of Gaussian (DoG) is used to find a set of key points in the image by generating the Gaussian Pyramid at the different octaves in Eq. (3).

$$Q(e,f,\lambda) = G(e,f,r\lambda) - G(e,f,\lambda) \tag{3}$$

where r is a constant scaling factor between levels in scale space. Afterward, the scale-invariant Laplacian of Gaussian (LoG) approximation is performed to generate a domain orientation for each key point. In the last stage, the feature descriptor is computed, capturing a concise depiction of the

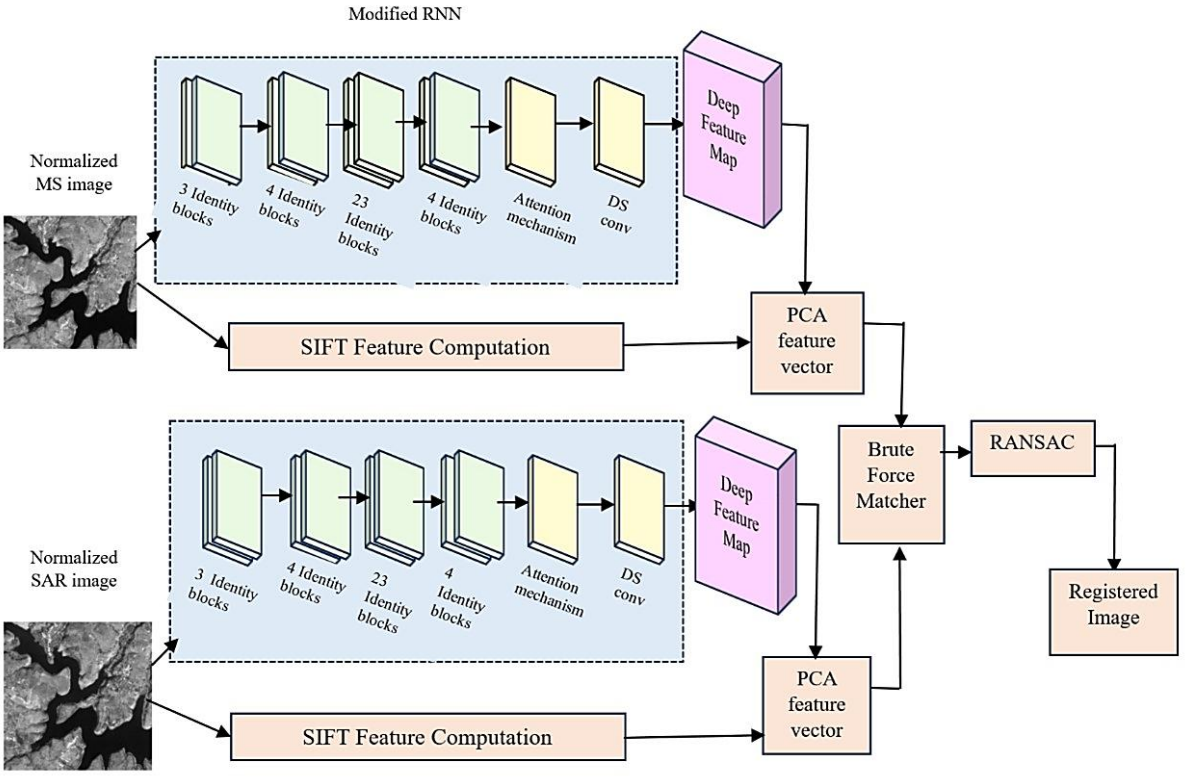

Fig. 1: Algorithm framework.

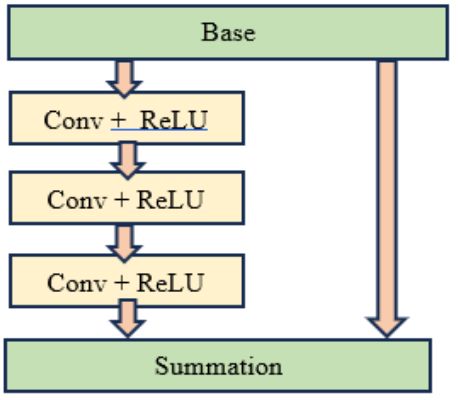

Fig. 2: Identity block.

local image region around each key point in Eq. (4).

$$Q(y) = Q_0 + \nabla Q^T y + \frac{1}{2} y^T H_Q y \quad (4)$$

where $y = (e, f, \lambda)^T$ is the offset vector, ∇Q is the gradient of Q, and HQ is the Hessian matrix of Q. These key points provide both orientation and key point location information, which is sufficient for basic image registration. Mathematical computation is done as per Eqs. (5) and (6).

Key point magnitude:

$$m(p,q) = \sqrt{\left(I_{(p+1,q)} - I_{(p-1,q)}\right)^2 + \left(I_{(p,q+1)} - I_{(p,q-1)}\right)^2} \quad (5)$$

Key point orientation:

$$b(x) = tan^{-1}\left(\frac{I_{(p,q+1)} - I_{(p,q-1)}}{I_{(p+1,q)} - I_{(p-1,q)}}\right) \quad (6)$$

where $I_{(p,q)}$ is the scale space representation of the given image, $m(p,q)$ is the magnitude of the chosen key point location, and $b(x)$ is orientation.

Therefore, SIFT has been proven to be a highly robust algorithm for local feature extraction due to its scale, rotation, and illumination invariance. It excels in complex scenes and multi-resolution tasks like image feature extraction and matching in satellite remote sensing imagery.

2.2 Network architecture

SIFT is proven to be one of the best algorithms for local feature extraction. However, this robustness in rotational and scale invariance is insufficient for achieving high-accuracy image registration. Nevertheless, incorporating Inception V3 and ResNet, as well as a generative adversarial network (GAN), can enhance image classification, registration, and prediction-related studies. A standard architecture of a CNN consists of a sequence of convolutional layers, pooling layers, and fully connected layers. These layers enable the extraction of various complex features. The increase in multiple convolutional layers leads to an increase in the number of parameters.

To reduce the number of network parameters, pooling is essential. This standard design network architecture is more suitable for classification or recognition applications. However, satellite images present challenging patterns to identify, diverse illumination ranges, large scale and resolution variances, which ultimately render it difficult to extract unique and accurate features using minimal computational cost. To

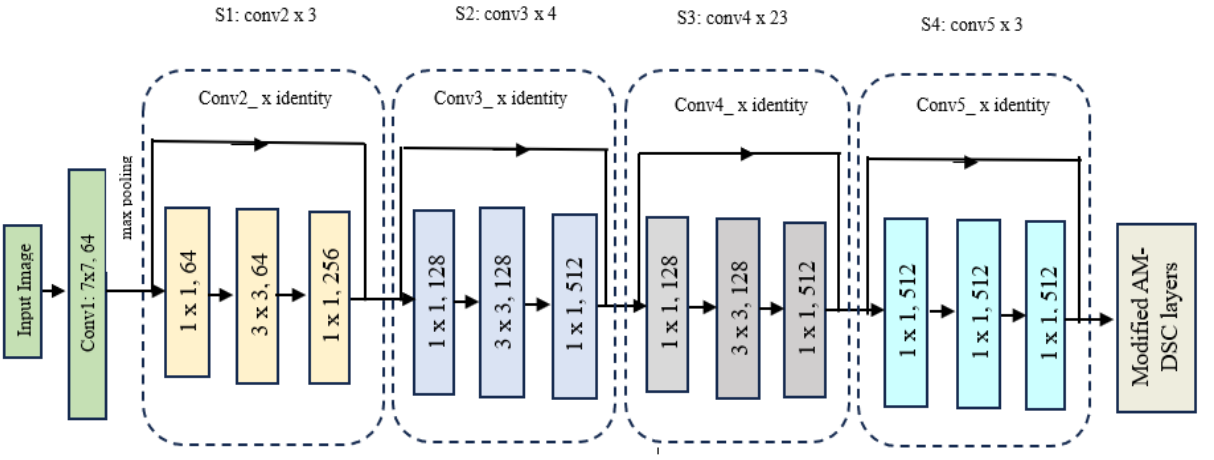

Fig. 3: Modified ResNet 101 network design.

resolve the issue, the proposed algorithm utilizes a modified RNN as shown in Fig. 3. The ResNet101 architecture, which claimed the top position in the 2015 ImageNet Large Scale Recognition Challenge, is distinguished by its shortcut connection that sums with the original input.[29] This feature sets it apart from conventional CNNs.

The operational sequence, referred to as the identity block, as depicted in Fig. 2, mitigates issues associated with increased model depth, such as vanishing gradients and elevated training errors. Furthermore, this model can be fine-tuned to focus on complex and diverse RS image features, enhancing its adaptability. Initially, zero padding is applied to ensure that the input tensor is compatible with the shortcut input size. Each subsequent step incorporates a convolutional layer (Conv), a batch normalization layer (BN), a Rectified Linear Unit (ReLU), and a bias. The residual learning can be mathematically represented as given in Eq. (7):

$$H(x) = x + F(x) \tag{7}$$

where x denotes the input image, $F(x)$ represents the output convolution operation, and the mapped output function is represented as $H(x)$, which helps the model to learn the input tensor more effectively. This residual learning capability enables the network to extract various-level features.

Fig. 3 depicts the proposed network model, which commences with input image processing in a convolution block comprising 64 filters of size 7×7 for convolution. ReLU activation is used to transform the negative values to zero and the positive values linearly proportional to the inputs. Further, the features are reduced using the max pooling layer. The residual block is then structured such that each block contains three layers (1×1, 3×3, and 1×1 convolutions). These are grouped into four stages. S1- 3 identity blocks (*i.e.*, 9 layers) S2- 4 identity blocks (*i.e.*, 12 layers) S3- 23 identity blocks (*i.e.*, 69 layers) and S4- 3 identity blocks (*i.e.*, 9 layers) as shown in Fig. 3. In place of the standard design's final layer, which typically consists of a global average pooling layer, an attention mechanism is integrated with depth-wise convolution is implemented as elaborated in the next point.

2.3 Attention mechanism

Descriptors are mathematical values that represent an edge, corner, line, point, or any feature present in the image. Basically, attention mechanisms are techniques designed to focus on the most significant regions of an image while disregarding irrelevant parts. The proposed algorithm used a neighborhood pattern descriptor operator for image parsing. This operator assigns codes to pixels based on comparisons with neighboring pixels, effectively capturing local texture information. The global average pooling operation is employed to aggregate the spatial features using the mathematical equation as follows in Eq. (8):

$$GAP_c = \frac{1}{h \times w} \sum_{i=0}^{h-1} \sum_{j=0}^{w-1} X_{i,j,c} \tag{8}$$

where h and w are the height and width of the feature map. $X_{i,j,c}$ is the activation value at pixel (i,j) of channel c. Furthermore, it establishes relationships at varying distances from the center pixel, enabling the capture of texture variation at different scales. Mathematical representation is as follows in Eq. (9):

$$GMP_c = \left(max_{i=0}^{h-1} \, max_{j=0}^{w-1} \, X_{i,j,c} \right) \tag{9}$$

This technique is characterized by its simplicity, robustness, and efficacy in identifying distinctive features from the image. These results are achieved through the operations delineated in Eq. (10).

$$A_c = \sigma(GAP_c + GMP_c) \tag{10}$$

where A_c is the attention map, $GAPc$ is the input feature map of global average pooling, $GMPc$ is the global max pooling feature map, and σ expresses a sigmoid activation function. This feature map is normalized using the sigmoid function. At the last stage feature map is reshaped for further use in Eq. (11).

$$X_{attend} = X * A_c \tag{11}$$

where X_{attend} is the final attention map while X denotes the input tensor.

2.4 Depth-wise separable convolution

To enhance the architecture's ability to extract robust image features, a depth-wise separable convolution kernel layer is incorporated. This technique involves the simultaneous application of a filter across all channels of the input image and performing element-wise multiplication of a filter with the overlapping regions of the input image. A depth-wise separable convolution is a sequential process that begins with a depth-wise convolution, where a single filter is applied to each input channel, followed by a pointwise convolution, which utilizes a 1×1 convolution to merge the output from the depth-wise convolution. This process can be represented as Eq. (12):

$$u = PointwiseConv\big(DepthwiaseConv(X)\big) \tag{12}$$

In depth-wise convolution, each input channel is independently convolved with a corresponding depth-wise filter as shown in Fig. 4. The relational operation is as follows: let $X \in R^{h\times w\times c}$ represent the input feature map, where h and w are dimensions of height and width, and c is the number of channels.

Let $K \in R^{k\times k\times c}$ denote the depth-wise convolution filter, with $k \times k$ as the spatial dimension. The resulting output of the depth-wise convolution $U \in R^{h\prime\times w\prime\times c}$ is computed in Eq. (13):

$$U_{i,j,c} = \sum_{m=0}^{k-1}\sum_{n=0}^{k-1} X_{i+m,j+n,c}\cdot K_{m,n,c} \tag{13}$$

where i and j iterate over the spatial dimensions of the output feature map, and c iterates over the output channels. The application of depth-wise convolution to each input channel is carried out independently, and subsequently, a pointwise convolution is used to integrate the output from the depth-wise convolution.

2.5 Feature fusion and selection

The next step involves feature fusion, wherein features extracted from the SIFT algorithm and the modified RNN network are concatenated. From this extensive feature set, it is essential to identify and retain only the most significant features. This necessitates dimensionality reduction to optimize the feature space while preserving the most informative and discriminative data aspects. For this purpose, PCA is employed, a widely recognized and effective technique for dimensionality reduction. PCA is initiated by computing the covariance matrix of normalized input features. This step prominently examines for pattern identification, dependencies, and the directions of maximum variance in the provided data. The covariance matrix encapsulates both individual feature variance and inter-feature relationships with their means. Covariance matrix C is computed using Eq. (14)

$$C = \frac{1}{n-1}Z^T Z \tag{14}$$

where Z is a matrix of data, Z^T is the transpose of Z, and n is the number of rows in the given matrix. From the covariance

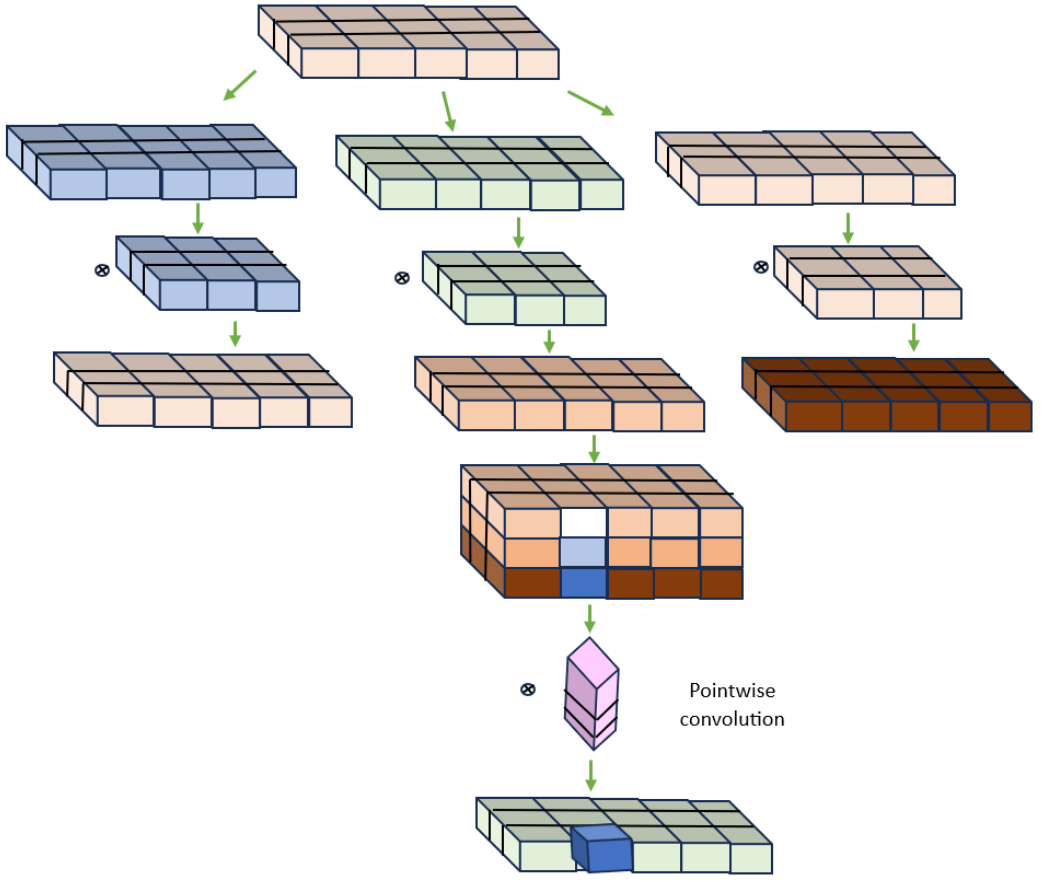

Fig. 4: Depth-wise separable convolution.

matrix, identify the principal components by calculating eigenvalues and eigenvectors using Eq. (15).

$$C¥_i = Ɛ ¥_i \quad (15)$$

where Ɛ is the eigenvalue that indicates the amount of variance along the principal component $¥_i$, $¥_i$ is the eigenvector to the corresponding eigenvalue. In closing sort, the matrix in descending order of eigenvalues and select the first p number of eigenvectors corresponding to top p eigenvalues.

2.6 Feature matching

The two sets of significant features are intensively processed through a brute force matcher (BFM). The BFM effectively compares the feature descriptors by determining the L2 Euclidean distance of each descriptor $d_1 \epsilon D_1$ from the source image to every descriptor d2 ϵ D2 of the reference image. The pair with the smallest distance is considered a match. All the matches are further processed for image transformation and reconstruction.

2.7 Image transformation and reconstruction

RANSAC is the most commonly used method for image transformation and reconstruction. This algorithm identifies and excludes incorrect matches by evaluating model consistency. Reliability, stability, and precision are pillars of this algorithm. Additionally, it demonstrates strong robustness against noise and errors in feature extraction, effectively eliminating mismatches. RANSAC identifies a mathematical model that fits a set of data points by iteratively testing subsets of data and filtering outliers. Unlike the polynomial transformation algorithm, this algorithm can also execute with fewer control points.

3. Experiments and discussion

3.1 Dataset and experimental settings

Every satellite image needs some preprocessing, as different sensors have different mechanisms to capture the image. The very first step is to choose the study area from the Earth's surface and prepare the shape file of that area. This shape file is utilized to extract the corresponding dataset from the available extensive datasets. It is necessary to identify the overlapping datasets of required sensors with the available dataset. The SAR-Sentinel 2 image dataset provides 12 bands, out of which three bands, Green, Blue, and Near Infrared (NIR), are combined using QGIS 3.18 version. The SAR dataset is obtained from the Earth Resources Observation and Science (EROS) Centre from the U. S. Geological Survey through QGIS tool. The other dataset of the optical-LISS III sensor is purchased from the National Remote Sensing Centre, Hyderabad, India (NRSC). This dataset comprises three bands, namely Green, Blue, and NIR. Both dataset band combination is performed using QGIS tool. SAR image has a 15-meter resolution, while the LISS III Multi-Spectral image has an 8-meter resolution.

In order to evaluate the robustness of the proposed algorithm, the publicly accessible OSDataset was employed for comprehensive experimentation. The SAR images are acquired from the Chinese GaoFen-3 (GF-3) satellite, which operates with Mult-polarised C-band capabilities. Utilizing the spotlight imaging mode, GF-3 is capable of producing geocoded images with a resolution of 1 meter and a ground coverage of 10 km. The optical images are sourced from the Google Earth platform and have been resampled to a resolution of 1 meter. The SAR images are unpacked before the registration process. The experimental setup utilized consists of a CPU Intel Core i7-12500H 12th gen 2.50 GHZ, 2GB NVIDIA GPU, 64-bit operating system Windows 11, Programming environment PyCharm 2024.3.

3.2 Performance assessment

Proposed algorithm performance is assessed by experimenting with ten SAR and multispectral image pairs. The compared methods include SIFT, ORB, Inception V3, VGG16, and ResNet50. Here, we utilize five evaluation indexes, RMSE and $RMSE_{LOO}$, CMR, PSNR, and SSIM, to evaluate the proposed algorithm. PSNR, RMSE, and SSIM measure the accuracy of the algorithm. $RMSE_{LOO}$ and CMR measure the robustness.

3.2.1 RMSE, $RMSE_{LOO}$

Let's consider that $I_{ref}(x_i', y_i')$ denotes the transformed coordinates of $I_{reg}(x_i, y_i)$ from the original image. RMSE is the square root of the average of the differences between these points, derived by Eq. (16).

$$RMSE = \sqrt{\frac{1}{N}\sum_{k=1}^{N}(I_{reg}(x_k, y_k) - I_{ref}(x_k', y_k'))^2} \quad (16)$$

where N represents the number of different pairs. I_{reg}, I_{ref} Indicate registered and reference images, respectively. $RMSE_{LOO}$ is computed as the mean of the summation of RMSE of N-1 points by leaving the kth pixel of both images, *i.e.*, RMSE by leave-one-out method, evaluated using Eq. (17).[28]

$$RMSE_{LOO} = \sqrt{\frac{1}{N}\sum_{k=1}^{N-1}(I_{reg}(x_k, y_k) - I_{ref}(x_k', y_k'))^2} \quad (17)$$

3.2.2 SSIM

The SSIM considers image contrast, image structural similarity, and image brightness. A large SSIM value implies better image reconstruction quality, determined using Eq. (18).

$$SSIM = \frac{(2\mu_{reg}\mu_{ref}+C_1)(2\sigma_{ref\,reg}+C_2)}{(\mu_{reg}^2+\mu_{ref}^2+C_1)(\sigma_{reg}^2+\sigma_{ref}^2+C_2)} \quad (18)$$

where μ_{reg}, μ_{ref} are the means of spectral values for the registered and reference image, respectively. $\sigma_{reg}^2, \sigma_{ref}^2$ are the variances of spectral values for both images, representing the spectral spread of energy. $\sigma_{ref\,reg}$ is the covariance between the two input images. C_1, C_2 are the stability constants.

3.2.3 PSNR

This is a crucial metric for measuring the quality of an image after processing. It evaluates the ratio of the maximum achievable power of an input signal to the power of the noise, defined using Eq. (19).

$$PSNR = 10 \cdot log_{10} \frac{255}{MSE} \quad (19)$$

where *MSE* denotes the mean square error of the registered image.

3.2.4 CMR

It quantifies how well the algorithm aligns the features from the reference image with the target image. A higher value of CMR indicates a more accurate registration. *CMR* computation can be done by utilizing Eq. (20).

$$CMR = \frac{Number\ of\ correct\ matches}{Total\ number\ of\ matches} \quad (20)$$

where *Number of correct matches* is nothing but the number of inliers, and *Total number of matches* is the number of total features extracted.

3.3 Comparative results and analysis

In the proposed method, for the first image pair, SIFT, *i.e.*, local 4485×1152 features and 1×1024 deep features extracted from the multispectral image, likewise, 961×1152 and 1×1024 are extracted from the SAR image. Through PCA, a significant portion of low-priority features are identified and excluded as outliers, which effectively reduces the number of features. These scaled-down features are 4438×224 and 3800×224 of multispectral and SAR images, respectively. The network layer details are as per Table 1. The network uses MSE as a loss function for training purposes. This feature reduction has been achieved to a significant degree with high efficiency and ensuring the most relevant and impactful attributes are retained. These parameters are tested using BFM for short-distance pairs. In the final stage, RANSAC ensures the reconstruction of a high-quality image by applying robust outlier rejection techniques and optimizing the transformation parameters for accurate positioning.

Comparative results of the proposed method with state-of-the-art methods for ten multispectral and SAR image pairs are given in Tables 2-4. For methods that introduced black borders in registered images, refer to Figs. 6-9, these borders were removed prior to calculating the quantitative parameters to ensure accurate calculations. Tables 2-4 demonstrate that the proposed feature fusion-based method outperforms the comparative methods, the value of CMR is significantly improved over the traditional SIFT and ORB, and the advanced transform radiation-variation insensitive transform (RIFT2)-based algorithms. RIFT2-based algorithm is implemented by Li *et al*.,[47] which is equivalent to the method implemented by Lian *et al*.[48] Proceeding to the subsequent significant parameter, traditional methods and the RIFT2 method have achieved a maximum accuracy of 25% in identifying correct pairs.

Table 1: Network layer details

Layers (type)	Output shape	Parameters
Input_layer_4 (InputLayer)	(None, 224, 224, 3)	0
Resnet101 (Functional)	(None, 7, 7, 2048)	4,26,05,504
Global_average_pooling2d_4 (GlobalAveragePooling2D)	(None, 2048)	0
Global_max_pooling2d_2 (GlobalMaxPooling2D)	(None, 2048)	0
Add_101 (Add)	(None, 2048)	0
Dense_2 (Dense)	(None, 1)	2,049
Reshape_2 (Reshape)	(None, 1, 1, 1)	0
Multiply_2 (Multiply)	(None, 7, 7, 2048)	0
Depthwise_conv2d_2 (DepthwiseConv2D)	(None, 7, 7, 2048)	20,480
Batch_normalization_316	(None, 7, 7, 2048)	8,192
Re_lu_304 (ReLU)	(None, 7, 7, 2048)	0
Conv2d_314 (Conv2D)	(None, 7, 7, 1024)	20,98,176
Batch_normalization_317	(None, 7, 7, 1024)	4,096
Re_lu_305 (ReLU)	(None, 7, 7, 1024)	0
Global_average_pooling2d_5 (GlobalAveragePooling2D)	(None, 1024)	0

In contrast, neural network (NN)-based methods have managed to achieve up to 34% accuracy. Notably, the proposed method attains the highest accuracy ratio of 45% for the dataset pair P1. The comparison of RMSE and $RMSE_{LOO}$ indicates that the proposed method performs more effectively in terms of pixel-level accuracy than both the transform-based and NN-based approaches. The lower the RMSE value, the better the quality of a reconstructed image. Proposed feature fusion method effectively improves the RMSE from 4.45 to 1.018 and $RMSE_{LOO}$ from 1.0564 to 0.8854. Analysis of the RIFT2 method's overall performance across all ten datasets indicates that RIFT2 occasionally demonstrates superior performance, it does not consistently achieve favourable results across diverse datasets.

Additionally, PSNR and SSIM values imply a substantial improvement over other considered result. These results are achievable due to local features and deep features obtained through the effective AM-DSC mechanism. The overall performance analysis for dataset pair P2 is graphically represented as shown in Fig. 5 with respect to SSIM, CMR, RMSE, and $RMSE_{LOO}$. Rows (a) and (b) in Figs. 6, 7 presents the multispectral and SAR image pairs from various regions proximal to Pune, Maharashtra, India, respectively. In these figures, red border boxes indicate non-overlapping areas. In Fig. 6, rows (c), (d), (e), and (f) illustrate the images registered using SIFT, ORB, RIFT2, and VGG16 methods, respectively. Image Dataset pair 1 exhibits a high illumination difference, resulting in unsuccessful registration for the ORB algorithm, as shown in the first image in row (d) in Fig. 5. It is observed in several non-overlapping image pairs that these methods are unable to successfully recover the non-overlapping regions, yielding poor quantitative parameter values.

Furthermore, particularly for image pair P1, the black borders are introduced, suggesting that these methods extract

Table 2: Comparative result analysis of dataset pairs 1, 2, 3, 4, 5 and OSdataset pair 1.

	Dataset Pair 1					Dataset Pair 2				
Method	RMSE ↓	$RMSE_{LOO}$ ↓	SSIM ↑	PSNR ↑	CMR ↑	RMSE ↓	$RMSE_{LOO}$ ↓	SSIM ↑	PSNR ↑	CMR ↑
SIFT	4.45	3.0564	0.35	13.25	0.23	4.01	3.115	0.41	18.75	0.15
ORB	-	-	-	-	-	4.28	3.355	0.48	19.55	0.21
RIFT2	1.05	1.047	0.42	19.98	0.28	1.32	1.009	0.85	23.78	0.20
VGG16	3.11	2.541	0.43	20.6	0.29	3.12	2.554	0.58	21.64	0.34
Inception v3	2.96	1.987	0.44	21.89	0.33	2.89	1.873	0.69	22.87	0.31
Resnet 50	1.99	1.095	0.47	21.75	0.36	1.87	1.0956	0.74	22.90	0.38
Proposed	1.018	0.8854	0.50	23.98	0.45	1.065	0.8859	0.98	23.92	0.44
	Dataset Pair 3					Dataset Pair 4				
Method	RMSE ↓	$RMSE_{LOO}$ ↓	SSIM ↑	PSNR ↑	CMR ↑	RMSE ↓	$RMSE_{LOO}$ ↓	SSIM ↑	PSNR ↑	CMR ↑
SIFT	3.84	2.95	0.33	16.98	0.24	3.54	2.87	0.39	19.87	0.29
ORB	3.98	2.45	0.39	16.11	0.29	3.78	2.92	0.29	18.22	0.31
RIFT2	1.25	0.9838	0.45	23.95	0.25	1.038	0.994	0.39	23.59	0.32
VGG16	3.44	1.95	0.38	19.87	0.34	2.87	1.79	0.41	20.66	0.32
Inception v3	2.74	1.65	0.48	21.99	0.24	2.15	1.55	0.38	20.45	0.35
Resnet 50	1.95	0.98	0.49	23.41	0.37	1.98	1.22	0.41	21.89	0.39
Proposed	1.091	0.8857	0.50	24.62	0.45	1.08	0.8857	0.50	23.72	0.45
	Dataset Pair 5					OSDataset Pair 1				
Method	RMSE ↓	$RMSE_{LOO}$ ↓	SSIM ↑	PSNR ↑	CMR ↑	RMSE ↓	$RMSE_{LOO}$ ↓	SSIM ↑	PSNR ↑	CMR ↑
SIFT	3.15	2.81	0.35	19.88	0.24	4.12	3.54	0.38	18.97	0.28
ORB	2.48	1.56	0.31	18.64	0.22	-	-	-	-	-
RIFT2	2.24	1.97	0.41	21.24	0.29	1.028	0.994	0.51	24.07	0.11
VGG16	2.99	1.77	0.47	20.44	0.37	3.12	2.12	0.41	20.97	0.29
Inception v3	3.45	2.32	0.41	19.99	0.29	3.46	1.98	0.47	22.43	0.34
Resnet 50	2.01	1.45	0.54	22.47	0.34	2.10	1.78	0.52	24.79	0.42
Proposed	1.067	0.8859	0.62	25.61	0.44	1.087	0.842	0.68	25.69	0.51

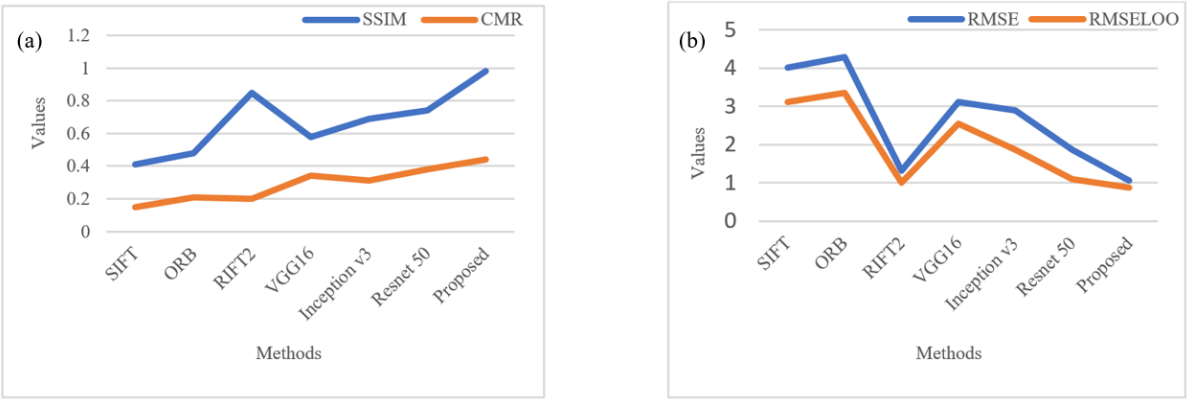

Fig. 5: Performance analysis for dataset pair P2 (a) SSIM and CMR, (b) RMSE and $RMSE_{LOO}$.

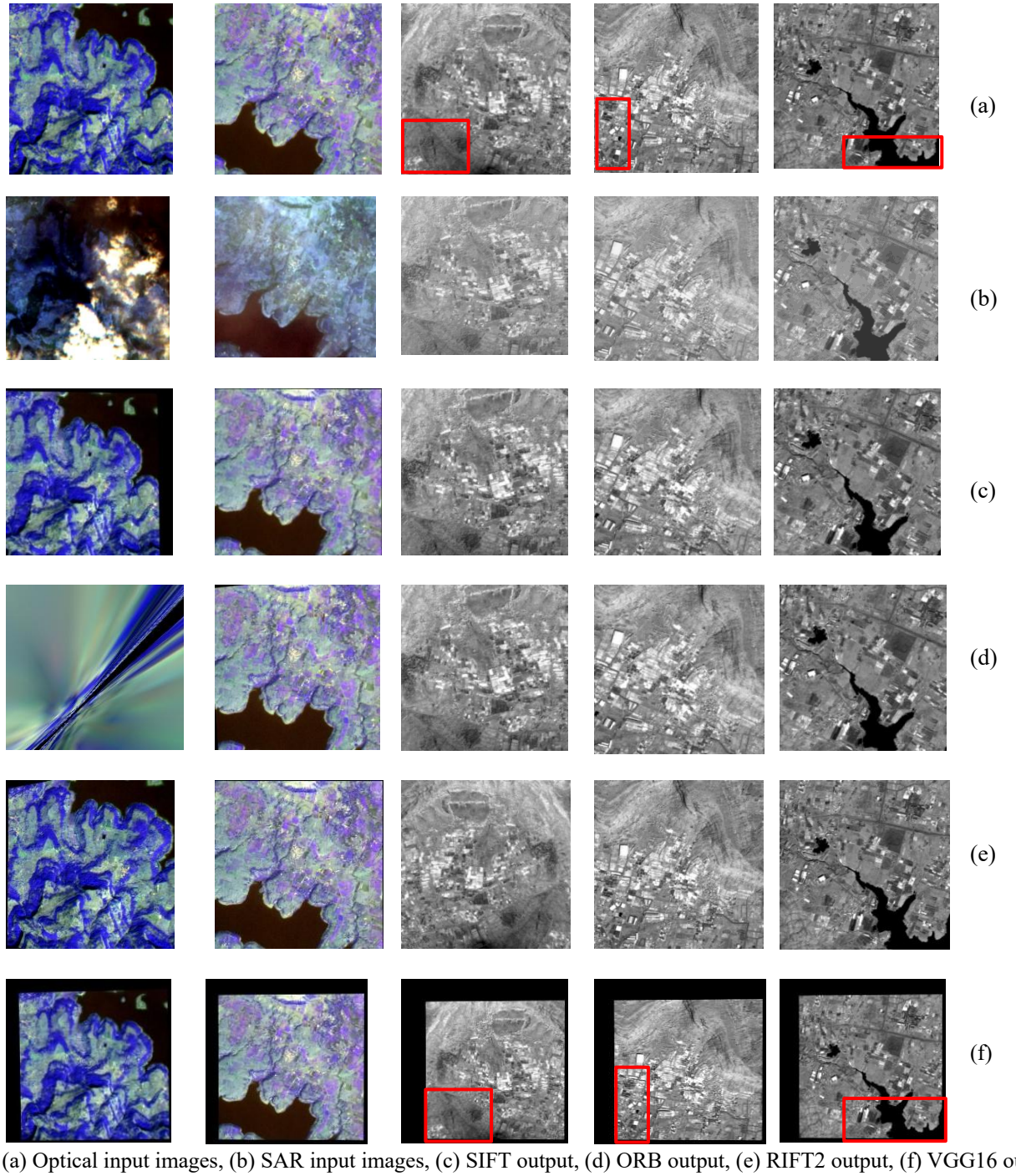

Fig. 6: (a) Optical input images, (b) SAR input images, (c) SIFT output, (d) ORB output, (e) RIFT2 output, (f) VGG16 output.

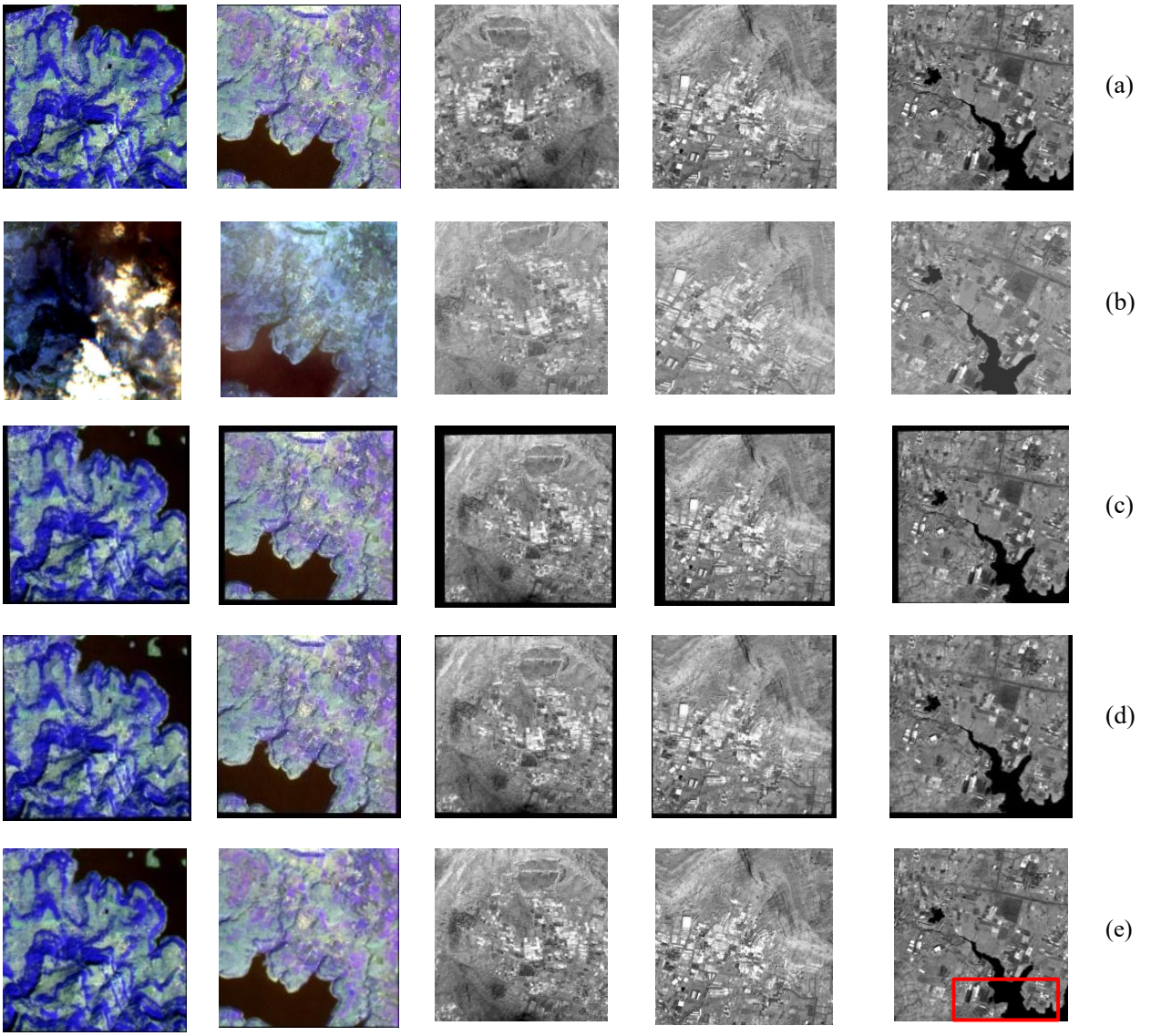

Fig. 7: (a) Optical input images, (b) SAR input images, (c) Inception V3 output, (d) ResNet 50 output, (e) Proposed method output.

Table 3: Comparative result analysis of OSdataset pairs 2 and 3.

	OSDataset Pair 2					OSDataset Pair 3				
Method	RMSE ↓	$RMSE_{LOO}$ ↓	SSIM ↑	PSNR ↑	CMR ↑	RMSE ↓	$RMSE_{LOO}$ ↓	SSIM ↑	PSNR ↑	CMR ↑
SIFT	3.39	2.114	0.27	15.89	0.18	3.98	2.24	0.24	19.97	0.24
ORB	-	-	-	-	-	-	-	-	-	-
RIFT2	1.23	0.9944	0.45	23.68	0.15	1.044	1.004	0.41	23.64	0.29
VGG16	2.89	1.99	0.38	18.47	0.19	3.012	2.231	0.35	20.85	0.45
Inception v3	2.05	1.456	0.41	19.78	0.25	2.671	1.987	0.4	21.79	0.54
Resnet 50	1.75	0.9859	0.52	22.87	0.33	1.574	0.954	0.41	22.86	0.78
Proposed	1.092	0.8839	0.6	24.05	0.45	1.092	0.8852	0.5	23.98	0.93

local features with low matching ratio. The red border box in different output images show the recovery of non-overlapping regions. Although NN-based methods show enhanced performance than traditional methods, redundant features persist, affecting the overall performance, as illustrated in Fig. 7. To validate the robustness of the proposed method, five additional dataset pairs from the OSdataset were assessed. These five OSdataset pairs are depicted in rows (a) and (b) of Figs. 8 and 9, respectively. The ORB method fails to register all the OSdataset pairs as shown in row (d) of Fig. 8, due to an insufficient number of correctly matching features. The performance of the comparative methods is largely consistent with that observed for previous dataset pairs, with the exception of RIFT2 method. RIFT2 method output for pairs 3, 4, and 5 are rotated, hence introducing black regions in the output image. Whereas, overall, the NN-based methods

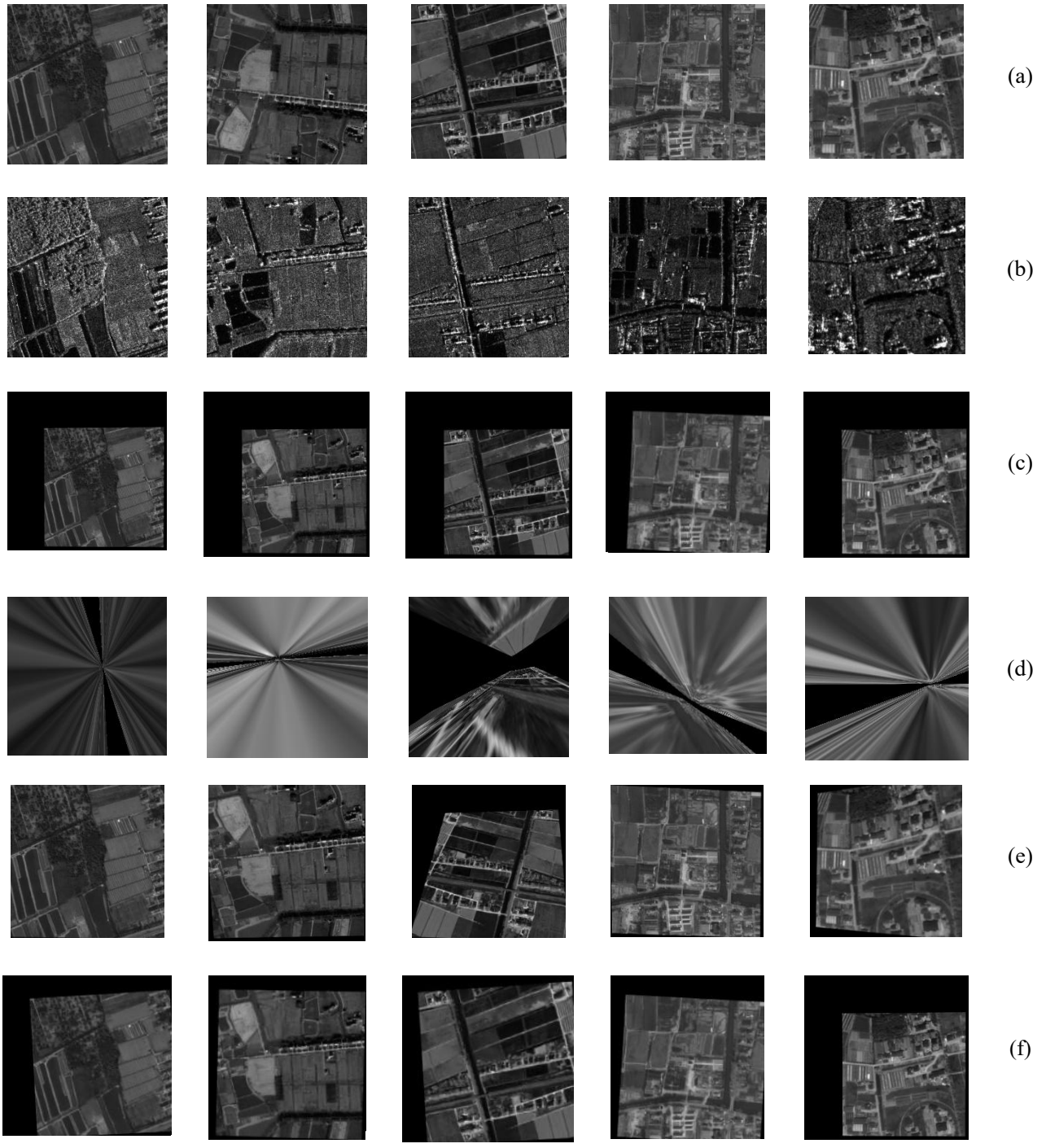

Fig. 8: (a) Optical input images, (b) SAR input images, (c) SIFT output, (d) ORB output, (e) RIFT2 output, (f) VGG16 output.

Table 4: Comparative result analysis of OSdataset pairs 4 and 5.

	OSDataset Pair 4					OSDataset Pair 5				
Method	RMSE ↓	$RMSE_{LOO}$ ↓	SSIM ↑	PSNR ↑	CMR ↑	RMSE ↓	$RMSE_{LOO}$ ↓	SSIM ↑	PSNR ↑	CMR ↑
SIFT	3.21	2.88	0.21	18.54	0.21	4.54	3.27	0.21	18.56	0.45
ORB	-	-	-	-	-	-	-	-	-	-
RIFT2	1.36	0.9962	0.362	23.11	0.31	1.42	0.959	0.4	23.88	0.25
VGG16	2.54	1.84	0.328	19.46	0.29	3.68	2.12	0.29	19.54	0.45
Inception v3	2.12	1.45	0.414	21.98	0.27	2.78	1.975	0.31	20.78	0.68
Resnet 50	1.65	0.987	0.596	22.47	0.39	1.67	0.978	0.42	21.96	0.79
Proposed	1.094	0.8836	0.628	23.67	0.45	1.094	0.8851	0.5	23.67	0.93

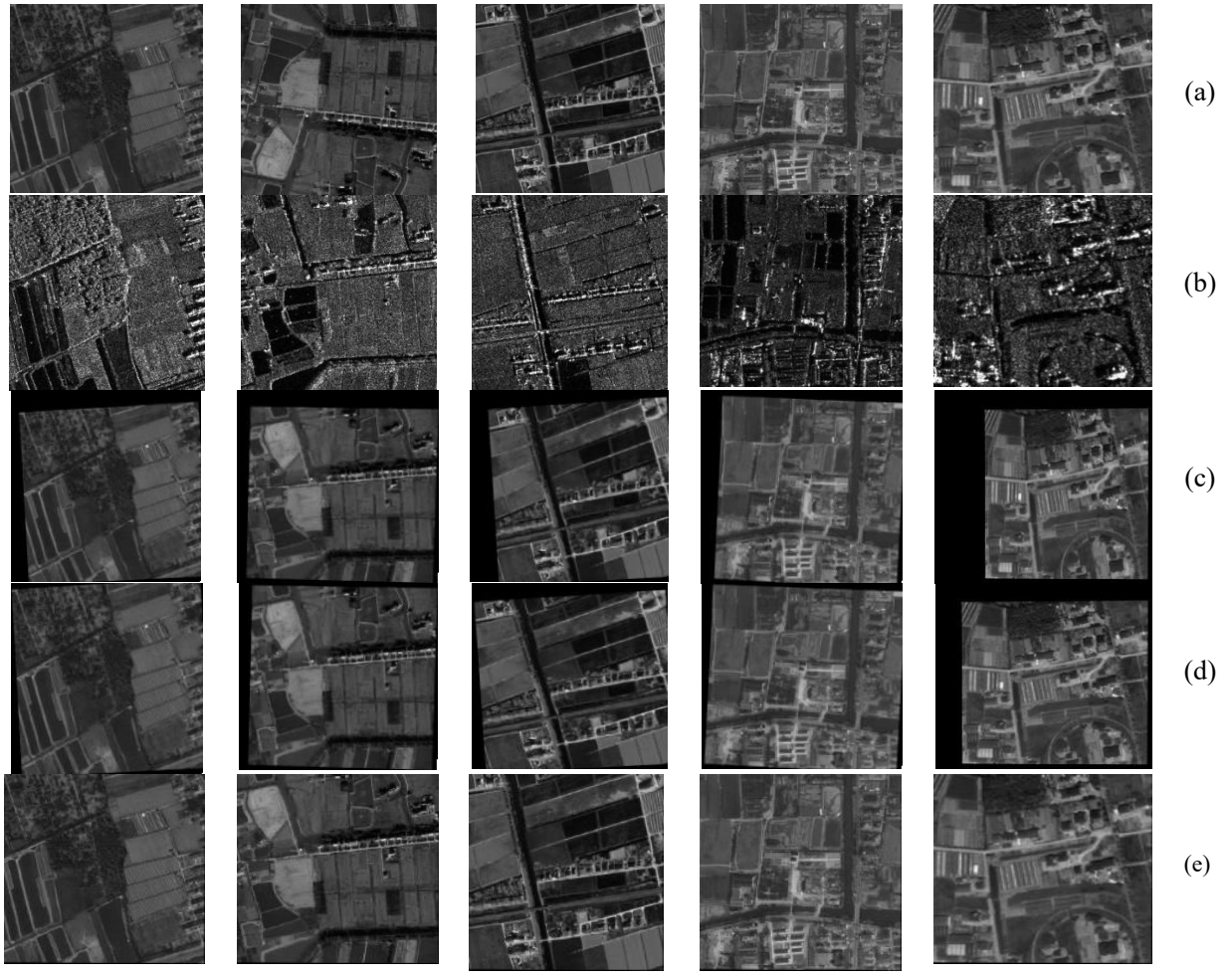

Fig. 9: (a) Optical input images, (b) SAR input images, (c) Inception V3 output, (d) ResNet 50 output, (e) Proposed method output.

achieve improved image recovery compared to the traditional methods, yet the introduction of black borders indicates the incorrectly mapped features. To overcome these issues, an RNN with AM-DSC is proposed. The traditional ResNet101 architecture can effectively capture the hierarchical features from the input image, but it may include redundant information, specifically for the image registration tasks. In contrast, the RNN with AM-DSC model emphasizes the accumulation of spatial information across the entire image while suppressing the noise and irrelevant features. Thus, this mechanism can be described as a filter to extract productive features.

4. Conclusion

In this research work, three ideas were proposed: 1) fusion of local and deep features using SIFT and a modified RNN model, 2) inclusion of an attention mechanism, and 3) tailoring of depth-wise separable convolutional kernel layer in RNN. This modified RNN network has significantly influenced the identification of the most accurate and important global features, which has subsequently improved the CMR. The experimentation exploited the five pairs of multispectral and SAR images, which exhibit notable variations in appearance, minor spatial displacements, and changes in illumination. The results clearly demonstrate that the proposed method outperforms existing approaches with respect to improved CMR and PSNR and minimized RMSE and $RMSE_{LOO}$. The evaluation metric effectively serves as a robust basis for validating the proposed algorithm's performance and reliability.

Conflict of Interest

There is no conflict of interest.

Supporting Information

Not applicable.

References

[1] L. Zeng, Y. Du, H. Lin, Jing Wang, J.Yin, J, Yang, A novel region-based image registration method for multisource remote sensing images via CNN, *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2020, **14**, 1821-1831, doi: 10.1109/JSTARS.2020.3047656.

[2] Y. Shi, L. Du, Y. Guo, Unsupervised domain adaptation for SAR target detection, *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2021, **14**, 6372-6385,

doi: 10.1109/JSTARS.2021.3089238.
[3] M. P. S. Tondewad, M. M. P. Dale, Remote sensing image registration methodology: review and discussion, *Procedia Computer Science*, 2020, **171**, 2390-2399, doi: 10.1016/j.procs.2020.04.259.
[4] A. K. Aggarwal, Fusion and enhancement techniques for processing of multispectral images, Unmanned Aerial Vehicle: Applications in Agriculture and Environment, Springer, Cham, 2019, 159-175, ISBN: 978-3-030-27156-5.
[5] H. Chen, H. Zhang, J. Du, B. Luo, Unified framework for the joint super-resolution and registration of multiangle multi/hyperspectral remote sensing images, *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2020, **13**, 2369-2384, doi: 10.1109/JSTARS.2020.2993629.
[6] C. Li, Y. You, J. Cao, W. Zhou, MoatNet: registration for multi-temporal optical remote sensing images using deep convolutional features, *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*, 2021, 2154-2157, doi: 10.1109/IGARSS47720.2021.9553197.
[7] L. Pallotta, G. Giunta, C. Clemente, SAR image registration in the presence of rotation and translation: a constrained least squares approach, *IEEE Geoscience and Remote Sensing Letters*, 2021, **18**, 1595-1599, doi: 10.1109/LGRS.2020.3005198.
[8] J. Zheng, W. Peng, Y. Wang, B. Zhai, Accelerated RANSAC for accurate image registration in aerial video surveillance, *IEEE Access*, 2021, **9**, 36775-36790, doi: 10.1109/ACCESS.2021.3061818.
[9] T. Dupuy, C. Beitone, J. Troccaz, S. Voros, 2D/3D deep registration along trajectories with spatiotemporal context: application to prostate biopsy navigation, *IEEE Transactions on Bio-Medical Engineering*, 2023, **70**, 2338-2349, doi: 10.1109/TBME.2023.3243436.
[10] H. Wen, Y. S. Xia, An improved SIFT operator-based image registration using cross-correlation information, *2011 4th International Congress on Image and Signal Processing*, October 15-17, Shanghai, China, IEEE, 2011, 869-873, doi: 10.1109/CISP.2011.6100362.
[11] S. Stempliuk, D. Menotti, Agriculture multispectral UAV image registration using salient features and mutual information, *IGARSS 2020 - 2020 IEEE International Geoscience and Remote Sensing Symposium*, September 26-October 2, Waikoloa, HI, USA, IEEE, 2020, doi: 10.1109/igarss39084.2020.9323325.
[12] Y. Chen, G. Liu, H. Chen, Multi-temporal remote sensing image registration based on multi-layer feature fusion of deep residual network, *2019 International Conference on Intelligent Informatics and Biomedical Sciences (ICIIBMS)*, November 21-24, 2019, Shanghai, China, IEEE, 2019, 363-367, doi: 10.1109/ICIIBMS46890.2019.8991506.
[13] X. Ma, W. Zhao, X. Hao, Y. Yang, K. Yang, Remote sensing image registration with adjustable threshold and variational mixture transformation, *IEEE Geoscience and Remote Sensing Letters*, 2020, **17**, 765-769, doi:10.1109/LGRS.2019.2936396.
[14] M. Zhang, Y. Yang, Q. Jiang, S. Zhang, A fast registration method based on line features, *2020 39th Chinese Control Conference (CCC)*, July 27-29, 2020, Shenyang, China, IEEE, 2020, 2918-2923, doi: 10.23919/CCC50068.2020.9189080.
[15] P. Li, Y. Pei, Y. Guo, G. Ma, T. Xu, H. Zha, Non-rigid 2D-3D registration using convolutional autoencoders, *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, April 3-7, Iowa City, IA, USA. IEEE, 2020, doi: 10.1109/isbi45749.2020.9098602.
[16] H. Goncalves, L. Corte-Real, J. A. Goncalves, Automatic image registration through image segmentation and SIFT, *IEEE Transactions on Geoscience and Remote Sensing*, 2011, vol. 49, pp. 2589-2600, doi: 10.1109/TGRS.2011.2109389.
[17] F. Dellinger, J. Delon, Y. Gousseau, J. Michel, F. Tupin, SAR-SIFT: A SIFT-like algorithm for applications on SAR images, *2012 IEEE International Geoscience and Remote Sensing Symposium*, 22-27 July, Munich, Germany, 2012, 3478-3481, doi: 10.1109/IGARSS.2012.6350671.
[18] M. T. Hossain, S. W. Teng, G. Lu, M. Lackmann, An enhancement to SIFT-based techniques for image registration, *2010 International Conference on Digital Image Computing: Techniques and Applications*, December 1-3, Sydney, NSW, Australia, IEEE, 2010, 166-171, doi: 10.1109/DICTA.2010.39.
[19] J. Fan, Y. Wu, F. Wang, Q. Zhang, G. Liao, M. Li, SAR image registration using phase congruency and nonlinear diffusion-based SIFT, *IEEE Geoscience and Remote Sensing Letters*, 2015, **12**, 562-566, doi: 10.1109/LGRS.2014.2351396.
[20] H. O. Velesaca, G. Bastidas, M. Rouhani, A. D. Sappa, Multimodal image registration techniques: a comprehensive survey, *Multimedia Tools and Applications*, 2024, **83**, pp. 63919-63947, doi: 10.1007/s11042-023-17991-2.
[21] A. Kumar, SURF feature descriptor for image analysis, *Imaging and Radiation Research*, 2024, **6**, 5643, doi: 10.24294/irr.v6i2.5643.
[22] T. Kumari, P. Syal, A. K. Aggarwal, V. Guleria, Hybrid image registration methods: A review, *International Journal of Advanced Trends in Computer Science and Engineering*, 2020, **9**, 1134-1142, doi: 10.30534/IJATCSE/2020/36922020.
[23] L. Sawkmie, S. Samal, B. K. Balabantaray, Automatic image registration using SIFT and AKAZE: an application for preservation of cultural heritage, *2023 OITS International Conference on Information Technology (OCIT)*, December 13-15, Raipur, India, IEEE, 2023, 150-155, doi: 10.1109/OCIT59427.2023.10431240.
[24] M. Gong, S. Zhao, L. Jiao, D. Tian, S. Wang, A novel coarse-to-fine scheme for automatic image registration based on SIFT and mutual information, *IEEE Transactions on Geoscience and Remote Sensing*, 2014, **52**, 4328-4338, doi: 10.1109/TGRS.2013.2281391.
[25] K. Simonyan A. Zisserman, Very Deep Convolutional Networks for Large-Scale Image Recognition, *3rd International Conference on Learning Representations (ICLR 2015)*, San Diego, CA, USA, May 7-9, 2015, 1-14, doi: 10.48550/arXiv.1409.1556.
[26] X. Hu, J. Yang, J. Yang, A CNN-based approach for lung 3D-CT registration, *IEEE Access*, 2020, **8**, 192835-192843, doi: 10.1016/j.media.2021.102139.

[27] H. Chang, Remote sensing image registration based upon extensive convolutional architecture with transfer learning and network pruning, *IEEE Transactions on Geoscience and Remote Sensing*, 2023, **61**, 5404416, doi:10.1109/TGRS.2023.3290243.
[28] Y. Liu, M. Lin, Y. Mo, Q. Wang, SAR–optical image matching using self-supervised detection and a transformer–CNN-based network, *IEEE Geoscience and Remote Sensing Letters*, 2024, **21**, 4002505, doi: 10.1109/LGRS.2024.3355472.
[29] Y. Ye, C. Yang, G. Gong, P. Yang, D. Quan, J. Li, Robust optical and SAR image matching using attention-enhanced structural features, *IEEE Transactions on Geoscience and Remote Sensing*, 2024, **62**, 5610212, doi: 10.1109/TGRS.2024.3366247.
[30] S. Wang, D. Quan, X. Liang, M. Ning, Y. Guo, L. Jiao, A deep learning framework for remote sensing image registration, *ISPRS Journal of Photogrammetry and Remote Sensing*, 2018, **145**, 148-164, doi: 10.1016/j.isprsjprs.2017.12.012.
[31] D. Quan, S. Wang, M. Ning, T. Xiong, L. Jiao, Using deep neural networks for synthetic aperture radar image registration, *2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, Beijing, China, July 10-15, 2016, 2799-2802, doi: 10.1109/IGARSS.2016.7729723.
[32] Y. Chen, J. Li, D. Wang, Remote sensing image registration based on attention and residual network, *2020 5th International Conference on Intelligent Informatics and Biomedical Sciences (ICIIBMS)*, Okinawa, Japan, November 18-20, 2020. doi: 10.1109/iciibms50712.2020.9336428.
[33] X. Zhao, H. Li, P. Wang, L. Jing, An image registration method using deep residual network features for multisource high-resolution remote sensing images, *Remote Sensing*, 2021, **13**, 3425, doi: 10.3390/rs13173425.
[34] D. Fan, Y. Ye, L. Pan, S. Yan, A remote sensing adapted image registration method based on SIFT and phase congruency, *2011 International Conference on Image Analysis and Signal Processing*, October 21-23, 2011, Wuhan, China, IEEE, 2011, 326-331, doi: 10.1109/IASP.2011.6109056.
[35] D. B. Nurseitov, G. Abdimanap, A. Abdallah, G. Sagatdinova, L. Balakay, T. Dedova, N. Rametov, A. Alimova, ROSID: remote sensing satellite data for oil spill detection on land, *Engineered Science*, 2024, **32**, 1348, doi10.30919/es1348.
[36] H. Li, X. Yue, L. Meng, Enhanced mechanisms of pooling and channel attention for deep learning feature maps, *PeerJ. Computer Science*, 2022, **8**, e1161, doi: 10.7717/peerj-cs.1161.
[37] Z. Han, N. Lv, T. Su, L. Cong, Z. Dou, F. Yang, W. Li, L. Zhao, C. Chen, A structure consistency generative adversarial network for SAR-optical image registration, *2024 IEEE International Conference on Smart Internet of Things (SmartIoT)*, Shenzhen, China, November 14-16, 2024, 197-203, doi: 10.1109/SmartIoT62235.2024.00037.
[38] R. Liu, H. Zhang, Optical and SAR image registration with deep reinforcement learning, *IGARSS 2024 - 2024 IEEE International Geoscience and Remote Sensing Symposium*, Athens, Greece, July 7-12, 2024, 7375-7377, doi: 10.1109/IGARSS53475.2024.10642618.
[39] H. Wang, A. Li, Q. Ye, X. Zhu, L. Song, Y. Ji, A coarse-to-fine heterologous registration method for Infrared-Visible images based on MDC and MSMA-SCW descriptors, *Optics and Lasers in Engineering*, 2025, **190**, 108955, doi: 10.1016/j.optlaseng.2025.108955.
[40] X. Hu, Y. Wu, Z. Li, Z. Yang, M. Li, Multifeature alignment and matching network for SAR and optical image registration, *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2025, **18**, 352-367, doi: 10.1109/jstars.2024.3492278.
[41] P. Wang, Y. Liu, X. Liang, D. Zhu, X. Gong, Y. Ye, H. F. Lee, B. Huang, CIRSM-net: a cyclic registration network for SAR and optical images, *IEEE Transactions on Geoscience and Remote Sensing*, 2025, **63**, 1-19, doi: 10.1109/tgrs.2025.3540258.
[42] Z. Han, N. Lv, T. Su, L. Cong, Z. Dou, F. Yang, W. Li, L. Zhao, C. Chen, A structure consistency generative adversarial network for SAR-optical image registration, *2024 IEEE International Conference on Smart Internet of Things (SmartIoT)*, November 14-16, Shenzhen, China, IEEE, 2024, 197-203, doi: 10.1109/SmartIoT62235.2024.00037.
[43] W. Xiong, M. Sun, H. Du, B. Xiong, C. Zhang, Q. Ou, Z. Rao, Cosine similarity template matching networks for optical and SAR image registration, *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2025, **18**, 813-827, doi: 10.1109/JSTARS.2024.3504555.
[44] H. R, P. K. N, A comprehensive study on deep learning techniques used for SAR and optical image registration, *2024 International Conference on Artificial Intelligence and Quantum Computation-Based Sensor Application (ICAIQSA)*, Nagpur, India, December 20-21, 2024, 1-6, doi: 10.1109/ICAIQSA64000.2024.10882455.
[45] J. Chen, Y. Liu, S. Wei, Z. Bian, S. Subramanian, A. Carass, J. L. Prince, Y. Du, A survey on deep learning in medical image registration: New technologies, uncertainty, evaluation metrics, and beyond, *Medical Image Analysis*, 2025, **100**, 103385, doi: 10.1016/j.media.2024.103385.
[46] J. Li, W. Yu, Z. Wang, J. Xie, X. Zhou, Y. Liu, Z. Yu, M. Li, Y. Wang, Low-dimensional multiscale fast SAR image registration method, *International Journal of Applied Earth Observation and Geoinformation*, 2024, **135**, 104266, doi: 10.1016/j.jag.2024.104266.
[47] J. Li, P. Shi, Q. Hu, Y. Zhang , RIFT2: speeding-up RIFT with a new rotation-invariance technique, *Journal of Latex Class Files*, 2021, **14**, 1-6, doi: 10.48550/arXiv.2303.00319.
[48] Z. Lian, S. Tang, J. Han, Y. Wu, M. Zhang, Z. Chen, L. Zhang, A multilevel point-matching algorithm based on hierarchical feature detection and description for SAR-to-optical image registration, *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2025, **18**, 7318-7333, doi: 10.1109/JSTARS.2025.3546224.