Electronic Research Archive

https://www.aimspress.com/journal/ERA

ERA, 33(4): 1875–1901.
DOI: 10.3934/era.2025085
Received: 10 December 2024
Revised: 04 March 2025
Accepted: 18 March 2025
Published: 02 April 2025

Research article

CoReFuNet: A coarse-to-fine registration and fusion network for typhoon intensity classification using multimodal satellite imagery

Zongsheng Zheng[1], Jia Du[1,*], Yuewei Zhang[2] and Xulong Wang[3]

[1] School of Information, Shanghai Ocean University, Shanghai 201306, China
[2] Guangzhou Meteorological Satellite Ground Station, Guangzhou 510650, China
[3] Shandong Provincial Institute of Land Space Data and Remote Sensing Technology, Jinan 250013, China

* **Correspondence:** Email: m220901554@st.shou.edu.cn.

Abstract: Typhoons cause significant damage to coastal and inland areas, making the accurate classification of typhoon cloud images essential for effective monitoring and forecasting. While integrating multimodal data from different satellites can improve classification accuracy, existing methods often rely on aligned images and fail to account for radiometric and structural differences, leading to performance degradation during image fusion. In registration methods designed to address this issue, two-stage approaches inaccurately estimate deformation fields, while one-stage methods typically overlook radiometric differences between typhoon cloud images. Additionally, fusion methods suffer from inherent noise accumulation and insufficient cross-modal feature utilization due to cascaded structures. To address these issues, this study proposed a coarse-to-fine registration and fusion network (CoReFuNet) that integrated a one-stage registration module with a cross-modal fusion module for multimodal typhoon cloud image classification. The registration module adopted a one-stage coarse-to-fine strategy, using cross-modal style alignment to address radiometric difference and global spatial registration by affine transformations to resolve positional differences. Bidirectional local feature refinement (BLFR) then ensured precise adjustments, facilitating fine registration by evaluating feature points in each image. Following registration, the fusion module employed a dual-branch alternating enhancement (DAE) approach, which reduced noise by learning cross-modal mapping relationships and applying feedback-based adjustments. Additionally, a cross-modal feature interaction (CMFI) module merged low-level, high-level, intra-modal, and intermodal features through a residual structure, minimizing modality differences and maximizing feature utilization. Experiments on the FY-HMW (Feng Yun-Himawari) dataset, constructed using data from the Feng Yun and

Himawari satellites, showed that the CoReFuNet outperformed existing registration methods (VoxelMorph and SIFT) and fusion methods (IFCNN and DenseFuse), achieving 84.34% accuracy and 87.16% G-mean on the FY test dataset, and 82.88% accuracy and 85.54% G-mean on the HMW test dataset. These results showed significant improvements, particularly in unaligned data scenarios, highlighting the potential for real-world typhoon monitoring and forecasting.

Keywords: cross modal; deep learning; dual-branch network; typhoon intensity classification; multimodal fusion

1. Introduction

Multimodal remote sensing data involves various types of remote sensing images from different satellite sensors across multiple spectral bands. Compared to mono-modal data, it provides various perspectives on ground feature information, which enriches the information content of remote sensing data [1,2]. In remote sensing land-cover classification and target detection, multimodal data enhances ground object extraction accuracy and effectively addresses the challenge of identifying complex objects that are difficult to discriminate with mono-modal data.

Multimodal learning aims to join data from various modalities by designing different fusion methods to overcome intermodal differences, thereby effectively utilizing the multimodal feature information [3,4]. However, differences in data acquisition methods across modalities create challenges in data collection. Additionally, many issues such as sensor data missing and corruption cause modal data mismatches, leading to difficulties in obtaining paired modal data [5]. Cross-modal learning effectively addresses these challenges by reducing the dependency on paired datasets while making full use of multimodal feature information [6].

To take advantage of the feature information from different modalities, some works have explored fusion methods to extract complementary information between modalities, such as DenseFuse [7], general image fusion framework based on convolutional neural network (IFCNN) [8], and U2Fusion [9]. Although these methods have shown promising results in infrared and visible image fusion tasks, they rely heavily on rigorously registered input image pairs. Besides, these methods seldom consider the inevitable impact of unaligned multimodal data, which can introduce structural distortions and artifacts in fusion results. However, multimodal images from different sensors often exhibit discrepancies in viewpoints, imaging principles, and spatial positions, resulting in inherent misalignment. Directly fusing unaligned multimodal image data can degrade fusion results, failing to support modeling tasks effectively. Therefore, fusion methods designed to rely on rigorously aligned multimodal data have limited practical application, as obtaining strictly aligned multimodal data is challenging. Thus, it is crucial to make image registration before image fusion.

Multimodal image registration aims to align two or more images of the same scene captured from different sensors, time periods, and viewpoints. This process mitigates challenges caused by local description differences or the absence of corresponding feature points due to image intensity and scale differences between different modalities. Failure to address these issues can lead to poor performance of multimodal image fusion in subsequent tasks such as classification and object detection [10].

Traditional registration methods can be divided into two types: area-based method, feature-based method and optical flow method [11,12]. Area-based registration methods optimize the transformation function parameters by similarity measurement until the optimal parameters are found. The representative methods include mutual information (MI) [13] and cross correlation (CC) [14]. However, existing area-based registration methods iterate slowly and are easily affected by image noise. Feature-based registration methods obtain coordinate mapping parameters by solving the relationships between feature pairs to establish associations for registration [15]. The classic methods include scale invariant feature transform (SIFT) [16] and radiation-variation insensitive feature transform (RIFT) [17]. While more robust than area-based registration methods, feature-based registration methods struggle with handling free-form deformation [18]. Similar to area-based methods, optical flow estimation utilizes intensity and gradient consistency constrains to compute pixel displacement for coordinate recalculation to achieve image registration. Typical methods include typical global approach [19] and [20] specialized designs for large-scale motion deformations. Although traditional registration methods can achieve superior results, they are less effective when dealing with complex multimodal data due to the modality-gap and the prevalence of texture and structural distortions [21]. Consequently, methods suitable for mono-modal registration often perform poorly in multimodal registration tasks.

Deep learning-based multimodal registration methods can be categorized into two types: two-stage registration methods and one-stage registration methods. Two-stage registration methods typically consist of two separate cascade stages: coarse registration and fine registration. Tang et al. [22] conducted cardiac orientation coarse registration and point cloud fine registration on myocardial perfusion imaging (MPI) and computed tomography angiograms (CTA) data, respectively. Xu et al. [23] first applied coarse registration to the original to-be-registered image pairs and then optimized boundary alignment through fine registration using the auxiliary gradient-space. RFNet (unsupervised network for mutually reinforcing multi-modal image registration and fusion) [24] used domain alignment and affine transformation to achieve coarse registration in the coarse stage, followed by fine registration based on fusion results. Similarly, Xu et al. [25] extended RFNet with contrastive learning to transform multimodal registration to mono-modal registration. They also established mutual feedback between registration and fusion in the fine registration stage.

However, the two-stage registration methods discussed above are independent of each other, lacking compactness, which limits their accuracy in directly estimating the target deformation field. To tackle this issue, UMF (unsupervised misaligned infrared and visible image fusion via cross-modality image generation and registration) [26] used the cross-modality perceptual style transfer network to transfer multimodal registration into mono-modal registration. This approach aimed to eliminate the differences between modalities and achieve the connection between coarse and fine registration in one stage from the perspective of mono-modal registration. Similarly, Wang et al. [27] refined the registration process based on UMF from a multi-scale perspective, combing coarse registration for dense deformation field fusion and fine registration for progressive feature refinement into one stage. Different from the above, Zhang et al. [28] designed a reconstructible mask in the coarse registration stage and employed spatial transformation for feature alignment in the fine registration stage. This approach improves registration ability while reducing the risk of mismatches. Li et al. [29] calculated the projection transformation matrix using eight displacement parameters to perform one-stage remote sensing image registration. Despite their advancements, these methods barely consider the issue of inconsistent distribution caused by style differences among different remote sensing images. Consequently, they struggle with image distortion and robustness when dealing with complex remote sensing images.

To fully utilize the feature information from different modalities, research in cross-modal fusion primarily focuses on noise suppression and intermodal difference reduction. To address cross-modal noise, Mai et al. [30] introduced the multimodal information bottleneck to filter noisy information while retaining maximum information relevant to correct prediction. Wang et al. [31] used a mutual information minimization module to reduce redundant information between modalities, enhancing object detection performance. Similarly, Cui et al. [32] applied a refinement-regularizer to reduce noisy information, which explored the information bottleneck to text-image tasks. Different from them, Chen et al. [33] introduced the channel attention mechanism in a simple cascade structure to filter modal noise and improve fusion sufficiency.

To reduce the differences between modalities, current cross-modal methods can be categorized into two types: spectrum-level and feature-level methods. In cross-modal spectral-level learning, Sun et al. [34] proposed a spectral resolution enhancement method (SREM). This method uses the spectral angle weighted minimum distance matching technique to convert original multispectral data into hyper-spectral data pixel by pixel, thereby utilizing the spectral resolution of hyper-spectral and the spatial resolution of original multispectral data, enhancing classification accuracy. A similar work was presented in [35], which investigated the effect of spectral enhancement on soil erosion by establishing spectral information links between different modalities through spectral mixture analysis. Yokoya et al. [36] designed a coupled nonnegative matrix factorization (CNMF) method to combine hyper-spectral and multispectral data for learning based on the unmixing method. This integration produces high spatial and spectral resolution data, improving classification performance. Mou et al. [37] used near-infrared, red, and green spectral data along with digital surface model (DSM) data as network inputs to learn feature information from multiple modalities. Ma et al. [38] employed a two-stage method to extract water bodies at the pixel level, integrating optical and synthetic aperture radar (SAR) images to improve water body classification accuracy.

In cross-modal feature-level learning, Hong et al. [39] established a common subspace where the original data was mapped by locally aligning the manifold structure of two modalities. This common subspace contains rich information from multiple modalities, which can effectively facilitate information exchange between modalities. Due to the lack of labelled data, Hong et al. [40] proposed a new semi-supervised cross-modal learning framework that utilizes learnable manifold alignment (LeMA) to create a graph learning structure, facilitating subspace-based information transfer between multiple modalities. Li et al. [41] addressed the influence of low-quality modal data on results by using cross-modal interactions between modalities, preserving the most valuable information at different scales through adaptive weighted features. Supported by knowledge distillation, Chen et al. [42] used auxiliary modalities to enhance target modality representations, while incorporating an additional term to reduce the intermodal information discrepancy. According to image generation, Hong et al. [43] also added a self-generated adversarial networks (GANs) module and a mutual-GANs module to semantic segmentation, effectively eliminating the gap between multimodalities and realizing more efficient feature information transfer. To improve classification accuracy, Hong et al. [44] designed a self-adversarial module, an interactive learning module, and a label propagation module to address the problem of semi-supervised transfer learning in remote sensing limited cross-modal data. This method facilitates feature transfer from hyper-spectral to multispectral or SAR data. In addition, Huang et al. [45] used a residual structure in multimodal hypergraph tasks, ensuring effective fusion of information from different hypergraphs. Xu and Mao [46] employed a residual structure to fuse image and text multimodal features, improving multimodal sentiment analysis accuracy. He et al. [47] designed a context enhancement module with the attention mechanism to learn feature information.

Although the image fusion methods described above gain some performance, they have the probability of failing to contain the maximum information that is necessary for correct prediction through both low-level and high-level, intra-modality and intermodality features. Besides, they may fail to effectively avoid interference from redundant information [48].

In the multimodal typhoon intensity estimation, Chen et al. [49] used both infrared data and microwave rain rates data as the input to estimate typhoon intensity. Similarly, Lee et al. [50] directly stacked infrared satellite cloud images from different bands to estimate typhoon intensity. However, the simple stacking method made limited use of the feature information from multimodal data. To tackle this issue, Zhang et al. [51] extracted infrared and water vapor images separately with a two-stream convolutional neural network, combining the feature information from these two modalities in the middle stage of feature extraction. Jiang et al. [52] improved typhoon intensity estimation by using a Kalman filter and attention mechanism with infrared cloud image data from four different channels. However, their method usually lacked interaction between intra- and intermodal, low-level and high-level multimodal feature information.

Typhoon satellite cloud images obtained from different satellites or sensors can complement each other in terms of spatial resolution, temporal resolution, and other aspects, effectively utilizing these advantages to provide more stable and precise analysis for typhoon forecasting and early warning. However, multimodal typhoon cloud images captured by different satellites often exhibit significant differences in radiometric and geometric features due to variations in sensors and imaging principles. These differences often lead to misalignment, which severely degrades fusion results and hinders accurate classification. Additionally, inherent noise, which is common in multimodal data, further complicates the fusion process.

To address these challenges, this research introduces CoReFuNet (a coarse-to-fine registration and fusion network for typhoon intensity classification using multimodal satellite imagery), a novel method that integrates data distribution alignment, spatial structure alignment, and geometric registration. CoReFuNet also investigates how to manage and suppress inherent noise during the fusion process and how to fully leverage both intra-modal and intermodal features. The method aims to leverage the complementary information from different modalities more effectively, thus improving classification accuracy and enhancing real-world applicability in typhoon forecasting and early warning.

Previous researches have often overlooked radiometric differences between images from different sensors during registration, considering only structural alignment, which leads to limited fusion outcomes. Furthermore, simple cascaded fusion methods tend to ignore inherent noise and fail to fully leverage complementary information across modalities. The CoReFuNet method addresses these limitations by integrating data distribution alignment, geometric registration, and noise reduction, providing a more comprehensive and effective solution for multimodal fusion in typhoon cloud image classification.

The major contributions of this paper are summarized as follows.

(1) We propose a one-stage coarse-to-fine multimodal registration method that effectively addresses the inherent radiometric and geometric differences in typhoon cloud images captured by different satellites. By combining cross-modal style alignment with global affine transformation, our method achieves both image distribution alignment and geometric structure alignment by computing modality differences and utilizing a multi-layer perceptron (MLP) to transform the data distribution. This advancement significantly enhances image fusion accuracy, directly tackling the challenge of data misalignment, —a common issue in real-world applications such as typhoon monitoring.

(2) In the fine registration stage, we introduce a bidirectional local feature refinement (BLFR) module to further refine the alignment. The BLFR module enhances the coarse registration module by emphasizing local feature differences. It computes feature information at each local position and dynamically assigns weights using Softmax. This is combined with an attention mechanism that focuses on key spatial regions, improving alignment in areas with fine details or misalignment. This enhancement facilitates more accurate subsequent fusion.

(3) To address cross-modal data noise, we designed the dual-branch alternating enhancement (DAE) module, which consists of two branches, each operating on a different modality. Using an attention mechanism to focus on key information, the Himawari (HMW) modality branch suppresses Feng Yun (FY) data noise while enhancing HMW modal features, and the FY modality branch suppresses HMW modal noise while enhancing FY modal features. This dual-branch approach not only reduces redundancy but also enhances feature quality across both the FY and HMW data modalities, significantly improving fusion performance.

(4) To further enhance the utilization of complementary information between modalities, we propose the cross-modal feature interaction (CMFI) module. By introducing a residual structure in later stages, the CMFI module fully leverages feature information across different levels (from low to high), as well as both intra-modal and intermodal features. This module overcomes the limitations of early-stage residual methods, which often fail to capture complementary information between modalities, thus improving the overall feature integration.

The rest of this paper is organized as follows: Section 2 introduces the proposed CoReFuNet framework in detail; Section 3 describes the experimental setup, evaluation metrics, experimental results, and ablation studies; and Section 4 concludes the paper and outlines future research directions.

2. Methodology

2.1. Problem formulation

Figure 1 provides an overview of the overall workflow of this study. As shown in Figure 2, the CoReFuNet proposed in this paper consists of three parts. The unaligned multimodal typhoon cloud image dataset is taken as input. First, a two-branch convolution neural network independently extracts multimodal feature information, which is then processed by the one-stage registration module. During the coarse registration stage, preliminary registration is performed based on image distributions and spatial distribution. The fine registration stage focuses on local feature alignment to further align the multimodal features. Finally, in the cross-modal fusion stage, the DAE module and CMFI module refine multimodal feature integration by reducing redundant information and improving the utilization of feature information across intra-modal and intermodal, as well as low-level and high-level features, thereby enhancing fusion quality and overall model performance.

2.2. Feature extraction

First, the FY cloud image and the corresponding HMW cloud image are fed into the feature extraction module, as illustrated in the feature extraction part in Figure 2. A dual-branch convolutional neural network is employed to extract feature information from the two different modal data, respectively. The two different modalities are denoted as $\mathbf{X}_1 \in \mathbb{R}^{m \times n \times c_1}$ and $\mathbf{X}_2 \in \mathbb{R}^{m \times n \times c_2}$. Here, m

and n represent the height and width, respectively, of the two images, and c_1 and c_2 refer to the number of channels of the two modalities. The detailed architecture of the dual-branch convolutional network is shown in Table 1. Each branch consists of convolutional layers, the rectified linear unit (ReLu) and the max-pooling layer to reduce the data variance and computation complexity. The obtained features f_{FY} and f_{HMW} are then fed into the subsequent registration stage.

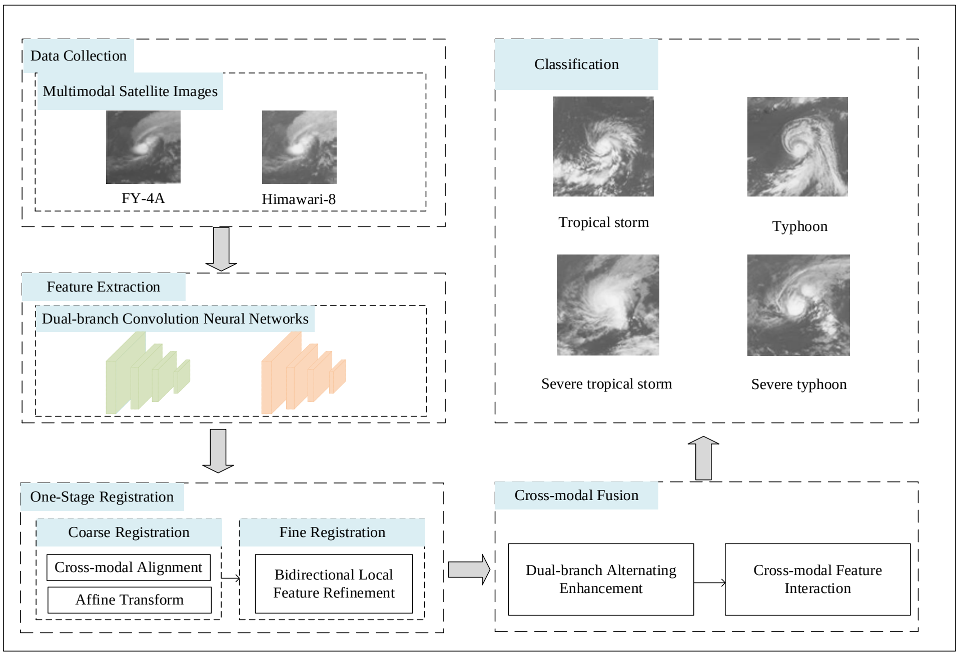


Figure 1. Overview of the CoReFuNet workflow for multimodal typhoon cloud image classification.

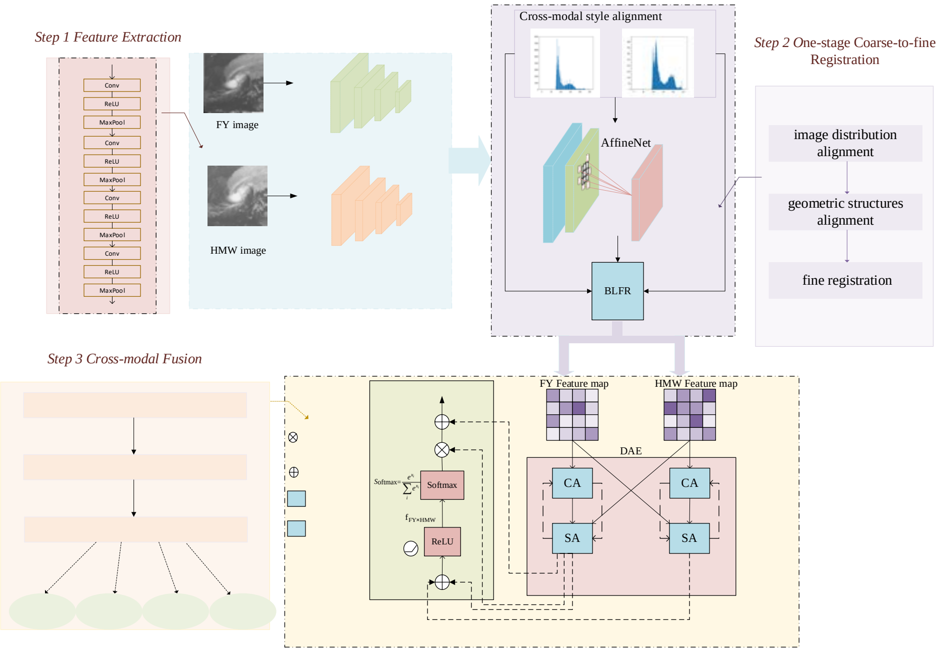

Step 1 Feature Extraction
Conv
ReLU
MaxPool
Conv
ReLU
MaxPool
Conv
ReLU
MaxPool
Conv
ReLU
MaxPool
FY image
HMW image
Cross-modal style alignment
AffineNet
BLFR
Step 2 One-stage Coarse-to-fine Registration
image distribution alignment
geometric structures alignment
fine registration
Step 3 Cross-modal Fusion
FY Feature map
HMW Feature map
DAE
CA
CA
SA
SA
Softmax
ReLU
$Softmax=\frac{e^{x_i}}{\sum_i e^{x_i}}$
$f_{FY\times HMW}$
⊗
⊕

Figure 2. An overview of the CoReFuNet architecture proposed in this paper. It consists of a feature extraction stage, a cross-modal register stage, and a cross-modal fusion stage.

Table 1. Convolutional network structure.

Conv1: 3 × 3 × 16	Conv2: 3 × 3 × 32	Conv3: 3 × 3 × 64	Conv4: 3 × 3 × 128
ReLu	ReLu	ReLu	ReLu
MaxPooling: 2 × 2	MaxPooling: 2 × 2	MaxPooling: 2 × 2	MaxPooling: 2 × 2

2.3. One-stage cross-modal register

2.3.1. Coarse registration

Multimodal images often exhibit significant differences in brightness and color, despite sharing modality-common information. During registration, this information may be mistaken as noise, leading to poor performance of traditional mono-modal registration methods when applied to multimodal registration. In previous multimodal registration tasks, many methods have considered reducing the large modality discrepancies during the registration process. Experimental results have shown that reducing the differences between modalities benefits multimodal registration. Therefore, as illustrated in Figure 3, cross-modal style alignment is designed to address these modality discrepancies by transforming the image distributions of FY data to match that of HMW while preserving essential structural features in the first part of the coarse registration stage. To achieve this, we calculate the mean and standard deviation of HMW data and use an MLP to determine optimal transformation parameters. These parameters are then applied to the normalized FY data to achieve a distribution shift, ensuring a more consistent representation modality. This step reduces style differences while retaining key images details, thereby improving subsequent registration accuracy. The process can be briefly described as:

$$\mathbf{f}_{FY\rightarrow HMW} = \mathrm{MLP}(\sigma_{\mathbf{f}_{HMW}})(\frac{\mathbf{f}_{FY} - \mu\mathbf{f}_{FY}}{\sigma_{FY}}) + \mathrm{MLP}(\mu_{\mathbf{f}_{HMW}}) \tag{1}$$

In the second stage of coarse registration, we employ deformable convolution to estimate affine transformation parameters in a data-driven manner. Unlike traditional methods that rely on fixed transformation matrices, our approach dynamically learns spatial offsets to achieve more effective multimodal alignment. Specifically, deformable convolution enables geometric structure alignment by adaptively adjusting the spatial positioning of FY modal data relative to HMW modal data. Unlike standard convolutions, which operate on fixed receptive fields, deformable convolution introduces learnable offset parameters that adapt dynamically during the convolution process. This allows the model to better capture complex spatial distortions and misalignments between FY and HMW data. By learning position offsets, the model achieves spatial alignment in a flexible and data-driven way, making it more robust to diverse transformations. This adaptive capability significantly enhances multimodal registration performance, particularly in scenarios where traditional affine transformations struggle with large geometric variations. As shown in Figure 2, the inputs to the deformable convolution are the result of cross-modal style alignment, denoted as $\mathbf{f}_{FY\rightarrow HMW}$, and the target alignment data, denoted as $\mathbf{HMW\_Conv}$. The offset of $\mathbf{f}_{FY\rightarrow HMW}$ with respect to $\mathbf{HMW\_Conv}$ is computed to align FY modal data with HMW modal data in spatial distribution. The affine transformation process using deformable convolution can be formulated as:

$$y(p)=\sum_{k=1}^{K} w_k \bullet x(p+p_k+\Delta p_k)\bullet \Delta m_k \tag{2}$$

where x represents the input features, y represents the output features, and k and K denote the index and the number of convolution kernels, respectively. p, $\mathrm{p_k}$, and $\Delta\mathrm{p_k}$ correspond to the center index, fixed offset, and learnable offset when using the k-th convolution kernel. Δm_k represents the modulation scalar at this position, and Δm_k denotes the weight.

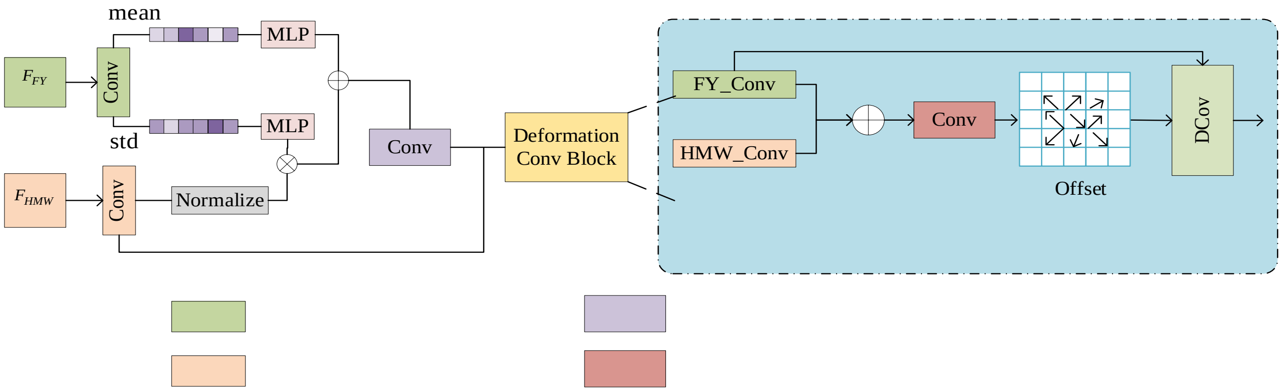


Figure 3. The architecture of the coarse registration module.

2.3.2. Fine registration

To further enhance registration results and improve the utilization of feature information, this paper introduces a BLFR module, which builds upon the coarse registration to refine alignment at a local level for further registration. As shown in Figure 4, the process involves two complementary branches: one for FY data and the other for HMW data. During the coarse registration, certain fine details from FY data may be lost. To solve this, the FY branch interacts the coarse registration results with the original FY data features. This interaction helps retain complementary information while refining alignment with HMW by computing similarity-based feature matching. Similarly, the HMW branch integrates the coarse registration results with HMW features, enhancing the structural information and refining feature alignment. These two branches calculate the offset for each feature position and dynamically assign weights through Softmax. Then, a spatial attention mechanism is applied to focus on the key spatial regions in the features, enabling more refined registration results. This step ensures that the registered FY data remains well-aligned with HMW while preserving detailed feature representations. By employing this bidirectional feature refinement, BLFR effectively mitigates registration errors caused by coarse misalignment and further improves multimodal feature fusion quality.

In the adaptive FY feature refinement branch, the feature vector of the FY data at spatial position (i,j) is denoted as $\mathbf{f}_{FY}^{ij}$, and $\mathrm{k}\times\mathrm{k}$ feature blocks at the same position are found in $\mathbf{f}_{FY\rightarrow HMW}$. Compute the feature information for each (i,j) and $(i+x,j+y)$ position, which is formulated as:

$$\mathbf{s}_{FY}^{i,j}=\frac{(\mathbf{f}_{FY-HMW}^{ij}(i+x,j+y))^{T}\,\mathbf{f}_{FY}^{ij}}{\|\mathbf{f}_{FY-HMW}^{ij}(i+x,j+y)\|_2\|\mathbf{f}_{FY}^{ij}\|_2} \tag{3}$$

Then SoftMax assigns different weights to the vectors in each position. Subsequently, refined feature information $\mathbf{out}_{FY}^{ij}$ for the enhanced FY data is obtained by applying SoftMax, convolution, and spatial attention mechanisms to the weighted features. This process retains FY feature information while aligning FY data with HMW data features.

$$\mathbf{w}_{FY}^{ij} = \mathrm{softmax}(\mathbf{s}_{FY}^{ij}) \tag{4}$$

$$\mathbf{out}_{FY}^{ij} = \mathrm{SA}(\mathrm{conv}(\mathbf{w}_{FY}^{ij})) * \mathrm{conv}(\mathbf{w}_{FY}^{ij}) \tag{5}$$

where $\mathrm{conv}(\bullet)$ denotes the convolutional layer and $\mathrm{SA}(\bullet)$ denotes the spatial attention mechanism.

In the adaptive HMW feature refinement branch, the coarsely registered feature $\mathbf{f}_{FY\to HMW}$ is used to enhance the HMW feature information. This involves using the feature vectors from HMW data at each spatial location (i, j) and the corresponding feature blocks in $\mathbf{f}_{FY\to HMW}$ to compute feature information $\mathbf{s}_{HMW}^{i,j}$ for each location. Finally, refined feature information $\mathbf{out}_{HMW}^{ij}$ for the enhanced HMW data is obtained, which can be expressed as:

$$\mathbf{out}_{HMW}^{ij} = \mathrm{SA}(\mathrm{conv}(\mathbf{w}_{HMW}^{ij})) * \mathrm{conv}(\mathrm{softmax}(\mathbf{s}_{HMW}^{ij})) \tag{6}$$

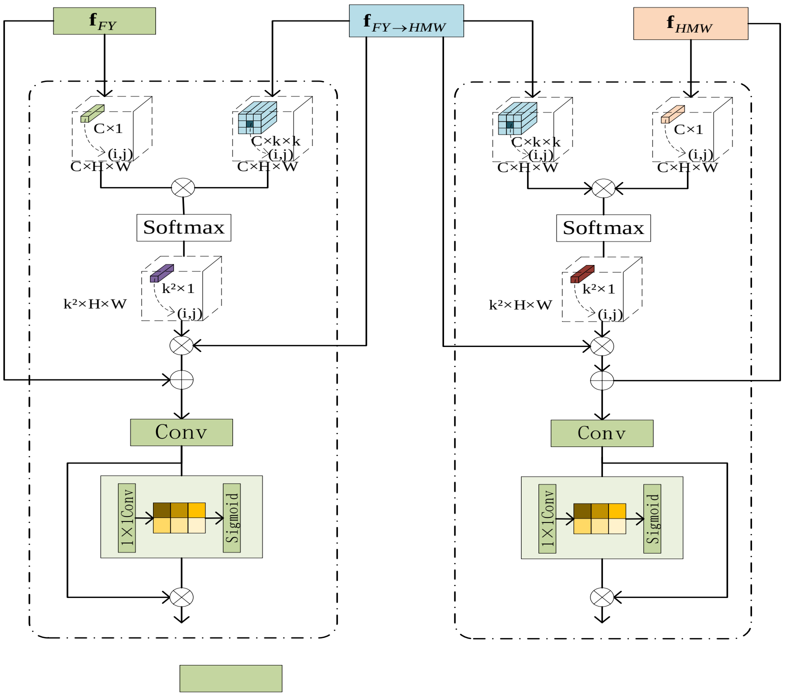


Figure 4. The architecture of bidirectional local feature refinement.

2.4. Cross-modal fusion

2.4.1. Dual-branch alternating enhancement module

Given the influence of inherent modal noise, the rich information provided by multimodal data can also amplify noise effects. If this noise is not addressed, directly fusing feature information from multiple modalities may exacerbate the noise impact and negatively affect the final result. To mitigate this issue, we propose the DAE module, which is designed to suppress noise and enhance feature information across modalities, as shown in Figure 5. The DAE module employs a dual-branch feedback structure, with one branch focusing on suppressing noise and enhancing features from the FY modality, and the other branch doing the same for the HMW modality. This feedback mechanism enables the two branches to interact with each other, effectively reducing noise while simultaneously enhancing important features. A key component of the DAE module is the bidirectional attention interactions, which plays a crucial role in learning and adjusting the relationships between the modalities. The attention mechanism allows the module to prioritize relevant features while filtering out nosy ones, ensuring that the fusion process is driven by informative, high-quality data. Additionally, the DAE module adapts to the specific characteristics of each modality by dynamically adjusting the attention weights, allowing for effective noise of suppression and enhanced feature representation. This dual-branch feedback structure, combined with the bidirectional attention mechanism, significantly reduces noise interference, improves feature representation, and ultimately enhances the fusion performance of multimodal data.

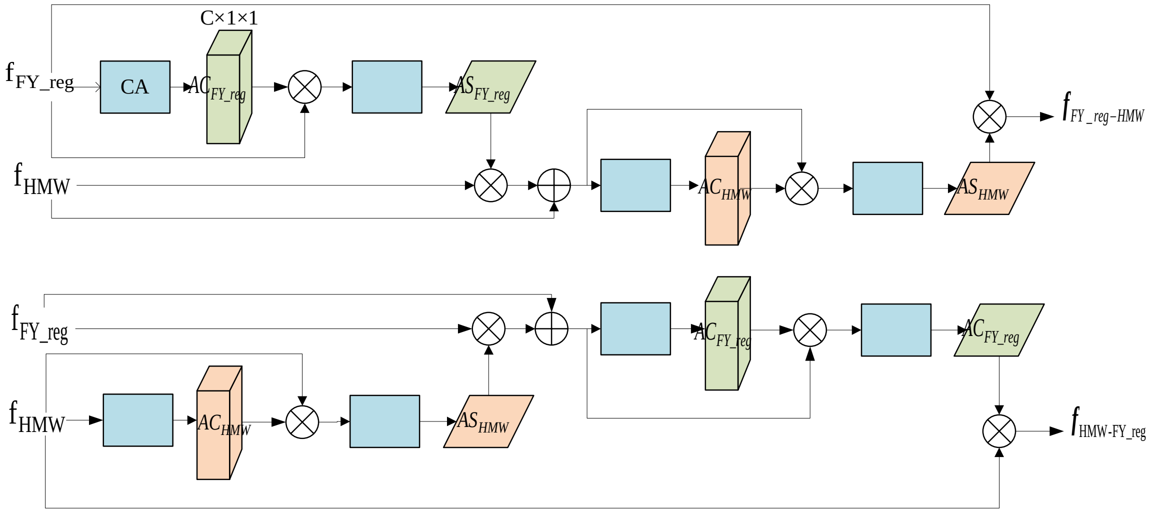


Figure 5. The architecture of DAE fusion module structure.

Taking the typhoon data as an example, the features obtained after the registration module are $\mathbf{f}_{FY\_reg}$ and $\mathbf{f}_{HMW}$, both of size $\mathbb{R}^{c\times h\times w}$. In fusion stage, feature $\mathbf{f}_{FY\_reg}$ is used to obtain the corresponding feature map $\mathbf{CA}_{FY\_reg} \in \mathbb{R}^{c\times 1\times 1}$ through the channel attention mechanism. Subsequently, feature $\mathbf{CA}_{FY\_reg}$ is used to obtain the feature map $\mathbf{SA}_{FY\_reg} \in \mathbb{R}^{1\times h\times w}$ through the spatial attention mechanism. The enhanced feature $\mathbf{f}_{FY\_reg\text{-}HMW}$ is calculated as follows:

$$\mathbf{f}_{FY\_reg\text{-}HMW} = \mathrm{SA}\left(\mathrm{CA}\left(\mathbf{f}_{FY\_reg}\right)\otimes\mathbf{f}_{FY\_reg}\right)\otimes\mathbf{f}_{HMW}\oplus\mathbf{f}_{HMW} \tag{7}$$

$$\mathbf{F}_{FY} = \mathrm{SA}\left(\mathrm{CA}\left(\mathbf{f}_{FY\_reg-HMW}\right)\otimes\mathbf{f}_{FY\_reg-HMW}\right)\otimes\mathbf{f}_{FY\_reg} \tag{8}$$

where $\otimes$ represents element-wise multiplication and $\oplus$ represents element-wise addition. $\mathbf{F}_{FY}$ is the feature after suppressing the HMW modal noise.

Similarly, for the HMW modal feature $\mathbf{f}_{HMW}$, $\mathbf{f}_{HMW\text{-}FY\_reg}$ and $\mathbf{F}_{HMW}$ can be obtained as follows:

$$\mathbf{f}_{HMW\text{-}FY\_reg} = \mathrm{SA}\left(\mathrm{CA}\left(\mathbf{f}_{HMW}\right) \otimes \mathbf{f}_{HMW}\right) \otimes \mathbf{f}_{FY\_reg} \oplus \mathbf{f}_{FY\_reg} \tag{9}$$

$$\mathbf{F}_{HMW} = \mathrm{SA}\left(\mathrm{CA}\left(\mathbf{f}_{HMW\text{-}FY\_reg}\right) \otimes \mathbf{f}_{HMW\text{-}FY\_reg}\right) \otimes \mathbf{f}_{HMW} \tag{10}$$

2.4.2. CMFI

To further reduce the influence of modal differences and improve the utilization of both low–level and high-level, intra-modal and intermodal features, we design a CMFI module. This module overcomes the limitations of cascade structure in existing modal fusion methods. By introducing residual connections in the later stages of the model, it reduces the impact of information loss and strengthens the feature representation ability. The combination of input and output features effectively enhances the model's ability to retain and optimize features from different modalities, improving feature utilization of both low-level and high-level and further enhancing the fusion of multimodal information. This approach retains modality-common information while obtaining modality-specific information at different levels from the two modalities. Consequently, the information between different modalities is effectively fused. As shown in Figure 2, the feature information $\mathbf{F}_{FY}$ and $\mathbf{F}_{HMW}$ obtained after the DAE module are first summed and passed through the ReLU (rectified linear unit) activation function to obtain $\mathbf{f}_{\mathrm{FY\times HMW}}$. Then the result $\mathbf{f}_{\mathrm{FY\times HMW}}$ is multiplied by $\mathbf{F}_{FY}$ after the SoftMax operation and finally summed with $\mathbf{F}_{HMW}$ to obtain the final output feature f. This output feature f contains both modality-specific information and modality-common information.

$$\mathbf{f}_{FY\times HMW} = \mathrm{ReLu}\left(\mathbf{F}_{FY} \oplus \mathbf{F}_{HMW}\right) \tag{11}$$

$$\mathbf{f} = \mathrm{softmax}\left(\mathbf{f}_{FY\times HMW}\right) \otimes \mathbf{F}_{FY} \oplus \mathbf{F}_{FY} \tag{12}$$

2.5. Cross-loss function

The loss function used in this paper consists of two components: the registration loss function and the classification loss function. For the registration loss function, mean squared error (MSE) is used to assess the alignment between FY features and HMW features, thus ensuring their alignment. In the loss function, y_i represents the predicted value of the i-th data and $\overset{\wedge}{\boldsymbol{y}}$ is the true value of the i-th data.

$$L_{\mathrm{align}} = \frac{1}{m}\sum_{i=1}^{m}(y_i - \hat{y})^2 \tag{13}$$

To comprehensively consider the influence of features from mono-modal and multimodal fusion in the classification loss function, let the given training set be denoted as $\left\{\left(\boldsymbol{x}_{FY}^{(i)}, \boldsymbol{x}_{HMW}^{(i)}, \boldsymbol{y}^{(i)}\right) | i = 1, 2, \cdots, N\right\}$, where N represents the number of training samples. Let $\boldsymbol{x}_{FY}^{(i)}$ and $\boldsymbol{x}_{HMW}^{(i)}$ represent the FY data and its corresponding HMW data of the i-th sample, respectively, and $\boldsymbol{y}^{(i)}$ represents the ground-truth for the i-th sample. During the feed-forward process of training, three output values will be obtained for each sample and the loss values are computed by the cross entropy loss function. The loss function during training is shown in Eq (14):

$$L = \lambda_1 L_{FY} + \lambda_2 L_{HMW} + \lambda_3 L_{\text{FY-HMW}} \tag{14}$$

where L_{FY} and L_{HMW} denote the loss functions for the FY and HMW modalities, respectively. When the output result is $\hat{\boldsymbol{y}}_{FY}^{(i)}$ and $\hat{\boldsymbol{y}}_{HMW}^{(i)}$ the true label is $\boldsymbol{y}^{(i)}$, respectively. $L_{FY\text{-}HMW}$ is designed to supervise the feature learning process after the fusion of the two modalities. λ_1, λ_2 and λ_3 represent the weight parameters for three different loss values, respectively. L_{FY} can be obtained as follows:

$$L_{FY} = -\frac{1}{N}\sum_{i=1}^{N}\left[\boldsymbol{y}^{(i)}\log\left(\hat{\boldsymbol{y}}_{FY}^{(i)}\right) + \left(1-\boldsymbol{y}^{(i)}\right)\log\left(1-\hat{\boldsymbol{y}}_{FY}^{(i)}\right)\right] \tag{15}$$

Similarly, we can obtain L_{HMW} and $L_{\text{FY-HMW}}$. The total loss function in this paper is:

$$L_{\text{total}} = \alpha L + \beta L_{align} \tag{16}$$

where α and β are the predefined nonnegative weight parameters. The weight parameters can be selected by conducting experiments with intervals of 0.1, evaluating performance on the validation set at each step.

3. Experimental analysis

3.1. Datasets

3.1.1. FY-HMW multimodal typhoon cloud image dataset

To validate the effectiveness of the proposed method, we constructed the FY-HMW multimodal typhoon cloud image dataset. The FY dataset consists of full disc 4KM L1 infrared images captured by the FY-4A satellite, obtained from the Wind and Cloud Remote Sensing Data Service Network (https://satellite.nsmc.org.cn/PortalSite/Data/Satellite.aspx). This dataset includes 14,056 typhoon images collected from March 2018 to September 2023. Correspondingly, the HMW dataset comprises infrared satellite cloud images acquired by the "Himawari1-8" satellite over the Northwest Pacific Ocean (https://agora.ex.nii.ac.jp/digital-typhoon/index.html.en). The typhoon intensity labels are provided by the National Institute of Informatics (NII), Japan. According to the international typhoon classification standard, the FY-HMW multimodal typhoon dataset is classified into four categories. To ensure fair evaluation, we randomly split the dataset into 80% for training, 10% for validation, and 10% for testing, as shown in Table 2. Figure 6 illustrates the comparison of multimodal typhoon data in each category.

3.1.2. RGB-NIR public scene dataset

The RGB-NIR (red (R), green (G), blue (B), and near-infrared) dataset [53] consists of 447 pairs of NIR and visible light images, captured from the same scene using an enhanced traditional digital single-lens reflex (DSLR) camera. These images are categorized into 9 distinct classes. Figure 7 presents a schematic illustrating representation of the images for each class, with the NIR image shown on the left and the corresponding visible light image displayed on the right. Due to the relatively small size of the dataset, we split it into a training set and a test set with a 9 : 1 ratio to better facilitate the training process. The number of training set samples for each category, following the order in Figure 7, is 41, 40, 42, 45, 44, 40, 39, 47, and 40, while the number of test set samples is 11 for each category.

Table 2. A benchmark for typhoon intensity established by NII and the number of training and test samples.

Intensity-based Typhoon classification	Maximum wind speed			Number of training sets	Number of test sets
	kt	m/s	Km/h		
Tropical storm	≥33~<48	≥17~<25	≥62~<89	4627	1157
Severe tropical storm	≥48~<64	≥25~<33	≥89~<118	2394	598
Typhoon	≥64~<85	≥33~<42	≥118~<150	2188	547
Severe typhoon	≥85	≥42	≥150	2037	509
Total				11245	2811

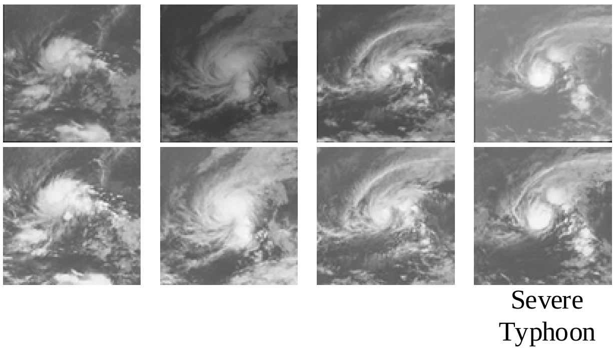
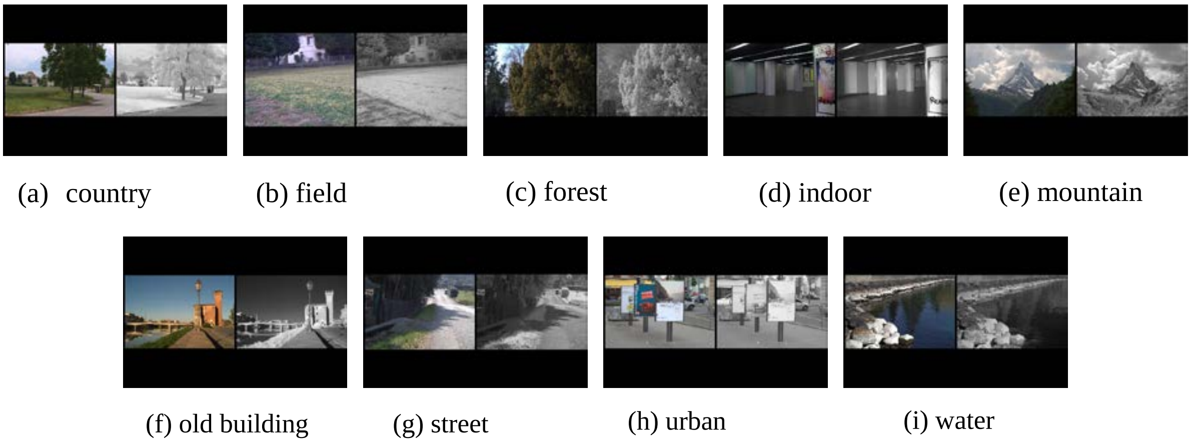


Figure 6. Comparison of multimodal typhoon data by category.


(a) country (b) field (c) forest (d) indoor (e) mountain

(f) old building (g) street (h) urban (i) water

Figure 7. Display of selected RGB-NIR datasets.

3.2. Implementation details

The proposed method in this paper is implemented using the PyTorch 1.12.0 framework and optimized with the Adam algorithm, employing a learning rate decay of 0.0004. Experiments are conducted on a Windows 10 system with an Intel Core i7-11700M CPU, 2.50 GHz processor, 16 GB RAM, and an NVIDIA GeForce RTX 3060 graphic card. The key hyperparameter settings used for training are summarized in Table 3 to facilitate reproducibility.

Table 3. Summary of hyperparameter settings used in training.

Hyperparameter	Value
Batch Size	16
Learning Rate	0.0001
Number of Epochs	100
Optimizer	Adam
Weight Decay	0.0004

The classification evaluation metrics for the proposed method include overall accuracy (OA), per-class accuracy, and geometric mean (G-mean). OA is the ratio of the number of correct predictions to the total number of correct predictions on all the test sets. G-mean is used in multiple classification tasks to evaluate the classifier's ability to correctly identify both positive and negative samples. It is particularly applicable for unbalanced samples among various categories. The value of G-mean is calculated using precision and recall.

$$\mathrm{OA} = \frac{\mathrm{M_c}}{\mathrm{M_t}} \tag{17}$$

$$\mathrm{G\text{-}mean} = \sqrt{\mathrm{Precision} * \mathrm{Recall}} \tag{18}$$

In Eq (16) $\mathrm{M_c}$ and $\mathrm{M_t}$ denote the number of correctly categorized samples and the total number of samples, respectively.

3.3. Ablation studies

To validate the effectiveness of the registration and fusion module proposed in this paper, ablation experiments were conducted on the FY-HMW dataset. Table 4 shows the test phase results for FY data and HMW data, respectively. Figure 8 illustrates the comparison of classification results for different modules on the test set.

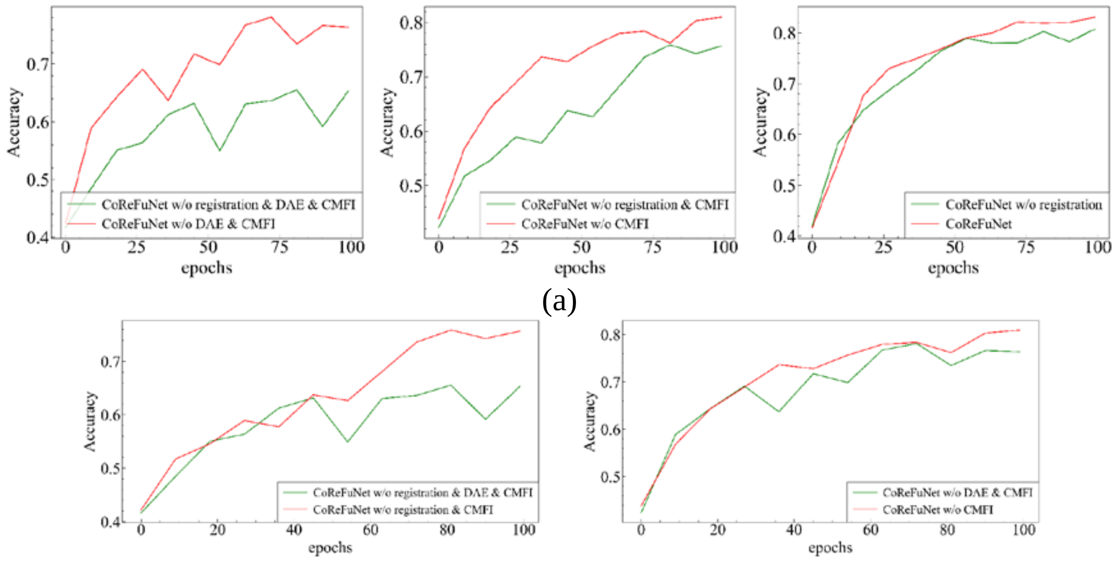
Effectiveness of one-stage coarse-to-fine registration: We evaluated the effectiveness of the one-stage coarse-to-fine registration module by disabling its core component. Figure 8(a) illustrates the improvements in classification accuracy when the coarse-to-fine registration module is applied, clearly demonstrating the impact of better alignment on model performance. Table 5 further quantifies these improvements, showing increased G-mean values that indicate the model's enhanced robustness after applying registration. Similar improvements were noted on the HMW test set. As shown in Table 4, incorporating the coarse-to-fine registration module significantly boosts performance by addressing both radiometric and geometric misalignments inherent in multimodal typhoon cloud images. This contribution is essential for achieving high-quality fusion and classification.

Table 4. Ablation study results on FY-HMW typhoon multimodal data in test phase for FY/HMW data.

Methods	Coarse-to- fine registration	DAE	CMFI	Tropical storm	Severe tropical storm	Typhoon	Severe typhoon	OA	G-mean
CoReFuNet w/o registration & DAE &CMFI				0.6146/0.8249	0.5726/0.6297	0.6685/0.7250	0.7405/0.8064	0.6381/0.7601	0.7477/0.8166
CoReFuNet w/o DAE &CMFI	✓			0.8927/0.9356	0.6493/0.6281	0.7326/0.6610	0.8164/0.7485	0.7957/0.7833	0.8405/0.8287
CoReFuNet w/o registration &CMFI		✓		0.9021/0.8979	0.6183/0.6183	0.5913/0.7589	0.7585/0.8004	0.7559/0.7932	0.7965/0.8363
CoReFuNet w/o CMFI	✓	✓		0.8918/0.9511	0.7243/0.6639	0.6573/0.7024	0.8563/**0.7764**	0.8046/0.8103	0.8448/0.8483
CoReFuNet w/o registration		✓	✓	0.9356/0.8807	0.6591/**0.7749**	0.7175/0.6591	0.7625/0.8024	0.8032/0.8018	0.8434/0.8434
CoReFuNet	✓	✓	✓	**0.9339/0.9545**	**0.7259/**0.7031	**0.7966/0.7778**	**0.8263/**0.7445	**0.8434/0.8288**	**0.8716/0.8554**

Table 5. Comparison of the proposed method with infrared/visible data in test stage on the RGB-NIR dataset by IFCNN and DenseFuse.

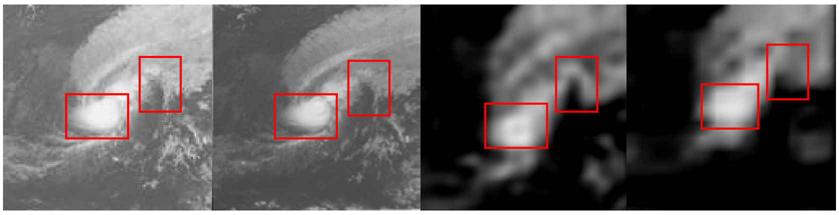
Methods	Country	Field	Forest	Indoor	Mountain	Old building	Street	Urban	Water	OA
IFCNN	0/0	0.4545/0.7273	0.9091/0.6364	0.8182/0.5455	0.5455/0.7273	0.0909/0.3636	0.5455/0.8182	0.2727/0.4545	0.1818/0.7273	0.4242/0.5556
DenseFuse	0/0.4545	0.5455/0.6364	0.9091/0.4545	0.9091/0.4545	0.4545/0.6364	0.0909/0.3636	0.8182/0.8182	0.6364/0.6364	0.5455/0.7273	0.5455/0.5758
PSFusion	0.3636/0	0.4545/0.6364	0.0909/0.0909	0.9091/0.5455	0.3636/0.1818	0.0909/0.2727	0.8182/0.1818	0.8182/0.7273	0.5455/0.1818	0.5657/0.3131
DAE+CMFI	0.3636/0.4545	0.8182/0.5455	0.9091/0.0909	0.7273/0.6364	0.2727/0.2727	0.4545/0.6364	0.9091/0.8182	0.6364/0.4545	0.6364/0.6364	**0.6398/0.6339**



(a)

(b)

Figure 8. Comparison of test results for the validation of the effectiveness of various modules for FY data: (a) Validation of the coarse-to-fine registration module, and (b) Validation of the DAE module.



(a) (b) (c) (d)

Figure 9. Visualization before and after adopting the coarse-to-fine registration module: (a) the original HMW image, (b) the original FY image, (c) fusion with registration, (d) fusion without registration.

To further assess the impact of the coarse-to-fine registration module on model performance, we visualized the registration process. Figure 9 compares a pair of images—one from the FY dataset and one from the HMW dataset—before and after applying the registration module. Notably, in the regions outlined in red, the typhoon eye contour and structural details become clearer, with more refined texture information after applying the registration module. This level of refinement is crucial for accurate typhoon cloud image classification.

Effectiveness of DAE: To evaluate the impact of the DAE module in the fusion stage, we replaced it with a simple concatenation method. The experimental results in Figure 8(b), consistently show that the DAE module improves classification accuracy, regardless of whether the registration module is included. Specifically, when the registration module is excluded, the DAE module increases classification accuracy by 0.1178 compared to the concatenation method on the FY dataset. Even when the registration module is included, the DAE module still provides a notable improvement of 0.0089 in overall accuracy. Furthermore, on the HMW dataset, the DAE module enhances both classification accuracy and G-mean. These results demonstrate that, compared to the simple concatenation method, the DAE module—incorporating an attention mechanism and bidirectional feedback adjustments—more effectively prioritizes important features across modalities, reduces noise interference, and improves overall classification performance. This is particularly significant in multimodal satellite imagery, where noise and redundancy often degrade classification accuracy.

Effectiveness of CMFI: To assess the effectiveness when the CMFI module is used in conjunction with DAE, we analyzed the experimental results, which indicate that the overall classification accuracy reaches its highest value on the FY dataset. This suggests that the CMFI module, with its residual structure, optimizes the utilization of both low-level and high-level features. The residual design ensures the preservation and effective propagation of critical information through the network, improving the model's ability to handle diverse features across different modalities.

Figure 10 presents the accuracy and error comparisons for different module combinations in the ablation experiments of CoReFuNet, evaluated across four categories of typhoon intensity cloud images. The incomplete versions of CoReFuNet exhibit a decline in both accuracy and error performance. In contrast, the complete CoReFuNet—incorporating the coarse-to-fine registration, DAE, and CMFI modules—achieves the most stable and accurate performance across all categories. The observed improvements across the test datasets underscore the importance of addressing both radiometric and geometric misalignments, reducing noise, and ensuring effective multimodal feature utilization in operational settings. By effectively handling multimodal satellite data, CoReFuNet enables more accurate classification of typhoon cloud images, which is critical for typhoon warning systems and disaster preparedness.

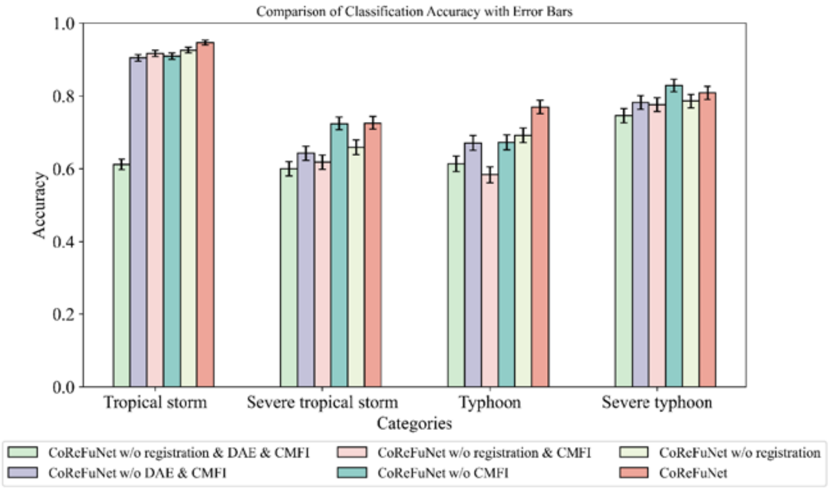


Figure 10. Ablation study comparing the performance of different CoReFuNet variants for typhoon classification.

3.4. Comparison with image fusion methods

Although the fusion method proposed in this paper is designed to fuse images after registration, it is also applicable to fully aligned multimodal image pairs. We conduct experiments on the RGB-NIR dataset, comparing the classification results of our method with those of IFCNN, DenseFuse, and PSFusion (practical infrared and visible image fusion network based on progressive semantic injection and scene fidelity constraints) [54] fusion methods. The results demonstrate that our proposed fusion method outperforms the others in terms of accuracy. Specifically, the combination of the DAE and CMFI modules achieves the highest accuracy on the test datasets, as shown in Table 5. These findings highlight that the DAE+CMFI fusion method effectively suppresses noise and maximizes the utilization of intra- and intermodal feature information, leading to superior performance in classification tasks on fully aligned multimodal image data.

3.5. Comparison with image registration + fusion methods

We use the combination of different registration and fusion methods as competitors to evaluate the overall performance of CoReFuNet on the FY-HMW dataset. The registration methods include VoxelMorph [55], a deep learning-based approach that has demonstrated strong performance in medical image registration, and SIFT, a widely used traditional registration method. Additionally, Bi-JROS [56], a recently proposed joint registration and optimization strategy, is included for comparison. The fusion methods are IFCNN, a general multimodal image fusion technique with broad applicability, and DenseFuse, which excels in preserving structural and feature information. Furthermore, we incorporate PSFusion, which emphasizes the representation of high-level semantic information during the fusion process, providing strong support for subsequent tasks. These methods represent a broad range of techniques, including traditional non-deep learning-based methods and deep learning-based methods. By comparing CoReFuNet with these established approaches, we ensure that both traditional and cutting-edge deep learning methods are considered. This allows for a more balanced and thorough assessment of CoReFuNet's strengths and its ability to address the challenges of multimodal data fusion and registration. By combining these methods, we conduct a comprehensive evaluation of CoReFuNet. The results, summarized in Table 7, are analyzed from two perspectives. To ensure a fair comparison, all methods were evaluated using the same input data formats and preprocessing steps. Additionally, hyperparameters were tuned consistently across all experiments to maintain uniform evaluation criteria. First, when using the same fusion methods, our one-stage coarse-to-fine registration module consistently outperforms both SIFT, VoxelMorph and Bi-JROS in classification accuracy across all fusion methods. This superior performance of our registration method can be attributed to its ability to align both data distributions and geometric structures. In contrast, VoxelMorph and Bi-JROS struggle with data distribution alignment, while SIFT has limitations in handling intermodal variability in multimodal registration tasks. By incorporating cross-modal style alignment, geometric structure alignment, and local feature refinement, our registration method significantly enhances the classification accuracy in multimodal typhoon data.

Having established the effectiveness of our one-stage coarse-to-fine registration module, we now examine how the DAE+CMFI fusion module further enhances classification performance by optimally integrating features from the aligned multimodal images. When the same registration methods are used, our DAE+CMFI fusion module consistently achieves higher classification accuracy compared to IFCNN, DenseFuse, and PSFusion. As shown in Table 6, when applying our proposed registration method, the classification accuracy on the FY dataset increases by 0.1281 compared to DenseFuse and by 0.0605 compared to IFCNN. Similarly, on the HMW dataset, the accuracy improves by 0.1213 over DenseFuse and 0.0363 over IFCNN. Notably, even when VoxelMorph is used for registration, our fusion module outperforms IFCNN, DenseFuse, and PSFusion in terms of classification accuracy. Although IFCNN is commonly used in multimodal applications, particularly in medical and remote sensing fields, it has limitations in fully leveraging multimodal feature information, which impacts its performance in this task. DenseFuse, by assigning weights to features during fusion using an L1 paradigm and SoftMax operation, improves feature utilization compared to IFCNN, resulting in better classification performance. Although PSFusion emphasizes the fusion of high-level semantic information, which benefits downstream tasks, it is sensitive to noise and its performance is limited when handling data from different sensors. However, the DAE+CMFI fusion method outperforms IFCNN, DenseFuse and PSFusion. The DAE module reduces noise by selectively enhancing important features, while the CMFI module maximizes intermodal feature utilization by introducing a residual structure that ensures effective integration of low- and high-level features from both modalities. This optimization of multimodal feature interaction leads to improved typhoon cloud image classification accuracy.

Table 6. Object evaluation results of different registration methods combined with fusion methods on the FY-HMW dataset in the test phase for FY/HMW data.

Registration Methods	Fusion Methods	Tropical storm	Severe tropical storm	Typhoon	Severe typhoon	OA	G-mean
Coarse-to-fine registration	IFCNN	0.8790/0.8755	0.5498/0.4649	0.5687/0.5951	0.6926/0.7325	**0.7153/0.7075**	**0.7672/0.7676**
VoxelMorph		0.8481/0.885	0.4780/0.3730	0.5405/0.5066	0.7764/0.6607	0.6964/0.6616	0.7639/0.7273
SIFT		0.9090/0.9305	0.4029/0.5122	0.4652/0.6177	0.6447/0.8244	0.6676/0.7612	0.7194/0.7928
Bi-JROS		0.8850/0.7803	0.5352/0.5090	0.5441/0.5857	0.7303/0.6687	0.7054/0.6644	0.7540/0.7398
Coarse-to-fine registration	DenseFuse	0.9296/0.9476	0.6052/0.6525	0.6987/0.6987	0.7485/0.7026	**0.7829/0.7925**	**0.8208/0.8301**
VoxelMorph		0.8764/0.9056	0.6020/0.4356	0.5725/0.5593	0.7924/0.6587	0.7441/0.6936	0.7935/0.7526
SIFT		0.8730/0.8773	0.5204/0.5449	0.5951/0.6911	0.6826/0.7565	0.7096/0.7480	0.7659/0.7882
Bi-JROS		0.8833/0.8249	0.6297/0.5024	0.6742/0.4953	0.8064/0.6747	0.7765/0.6655	0.8135/0.7384
Coarse-to-fine registration	PSFusion	0.9416/0.8996	0.6558/0.6020	0.6723/0.6911	0.7944/0.7984	**0.8021/0.7769**	**0.8361/0.8261**
VoxelMorph		0.8687/0.8361	0.5775/0.4470	0.6215/0.4972	0.7764/0.5908	0.7420/0.6434	0.8013/0.7142
SIFT		0.8412/0.8258	0.6248/0.5677	0.4256/0.4237	0.7146/0.6747	0.6929/0.6665	0.7515/0.7414
Bi-JROS		0.8927/0.8283	0.6558/0.5188	0.6930/0.5951	0.8024/0.6068	0.7890/0.6772	0.8278/0.7521
Coarse-to-fine registration	DAE+CMFI	0.9339/0.9545	0.7259/0.7031	0.7966/0.7778	0.8263/0.7445	**0.8434/0.8288**	**0.8716/0.8554**
VoxelMorph		0.9150/0.8747	0.6297/0.4927	0.6930/0.5857	0.7884/0.6806	0.7883/0.7021	0.8358/0.7621
SIFT		0.9107/0.9202	0.6117/0.5644	0.6761/0.7401	0.7864/0.7625	0.7790/0.7804	0.8200/0.8237
Bi-JROS		0.9073/0.8901	0.6150/0.4356	0.7213/0.6855	0.8204/0.6148	0.7929/0.7032	0.8359/0.7601

We also evaluated and compared the time cost and parameter count of different methods, as shown in Table 7. SIFT, a traditional non-deep learning-based method, exhibits the lowest time cost and parameter count, with most of its cost arising from the convolutional neural network used for feature extraction. Despite its minimal computational cost, SIFT’s performance in multimodal deep learning tasks is limited, resulting in the lowest classification accuracy. Similarly, IFCNN, a general-purpose fusion method with a simple structure, has been successfully applied in various fields. However, its capabilities are limited in complex multimodal tasks. Although it also shows low time and parameter usage, its performance lags behind in such tasks. In contrast, while CoReFuNet has a slightly higher time cost and parameter count compared to other methods, it significantly improves classification results. The increase in computational overhead is relatively small when compared to the substantial improvement in accuracy, demonstrating that the method is effective in addressing the challenges posed by complex multimodal classification tasks.

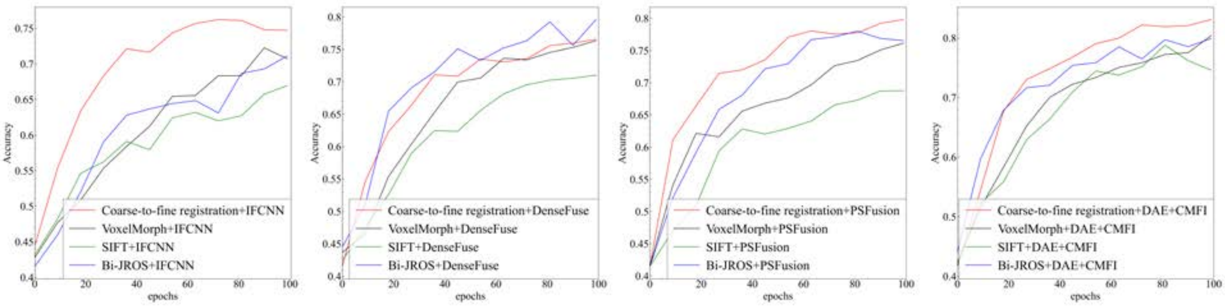


Figure 11. Comparison of our registration method with VoxelMorph, SIFT and Bi-JROS using the same fusion method.

Table 7. Comparison of time cost, flop, and parameters across different methods.

Registration Methods	Fusion Methods	Training Time (h)	Inference Time (s)	FLOPs (G)	Parameter Count (M)
Coarse-to-fine registration	IFCNN	3.27	37	0.535	16.100
VoxelMorph		3.22	24	5.174	13.390
SIFT		3.10	20	0.504	13.090
Bi-JROS		3.10	20	1.056	13.080
Coarse-to-fine registration	DenseFuse	3.27	37	0.462	16.100
VoxelMorph		3.22	24	5.100	13.390
SIFT		3.10	20	0.430	13.080
Bi-JROS		3.19	21	0.983	13.085
Coarse-to-fine registration	PSFusion	3.30	38	0.692	17.280
VoxelMorph		3.22	24	5.330	14.570
SIFT		3.19	22	0.660	14.260
Bi-JROS		3.27	37	1.213	14.264
Coarse-to-fine registration	DAE+CMFI	3.27	38	1.630	16.430
VoxelMorph		3.27	38	6.270	13.720
SIFT		3.27	37	1.600	13.410
Bi-JROS		3.27	38	2.153	13.415

We further analyze the qualitative results presented in Figures 12 and 13, which show the Grad-CAM (gradient-weighted class activation mapping) [57] heatmap visualizations for different registration and fusion methods across four typhoon categories. The four typhoon classes represent different intensity levels, and our method consistently outperforms competitors in classifying both weak and strong typhoons. The visualizations reveal that, under the same registration method, our fusion approach more effectively suppresses background noise and reduces interference from non-typhoon cloud systems. This reduction in background cloud interference significantly enhances classification accuracy. Additionally, our method prioritizes key typhoon features—such as the typhoon eye, eyewall, and spiral rainband—while minimizing attention to background clouds and nonrelevant structures. These findings align with the Dvorak technique [58], which assesses typhoon intensity based on structural features in typhoon cloud images. Among the compared fusion methods, PSFusion performs the second best, while IFCNN exhibits the lowest performance, which aligns with the quantitative classification results. When using the same fusion method, our one-stage coarse-to-fine registration method is particularly effective in reducing misclassification caused by pixel shifts, a common issue resulting from misalignments between modalities. These misalignments can cause typhoon features to be misplaced or misclassified. While VoxelMorph and Bi-JROS are effective for geometric alignment, it overlooks the importance of radiometric differences, occasionally leading to misclassification of non-typhoon structures. The SIFT registration method, with its limited ability to handle multimodal registration tasks, results in higher misjudgment and insufficient attention to crucial typhoon structures, making it more susceptible to background cloud interference during classification.

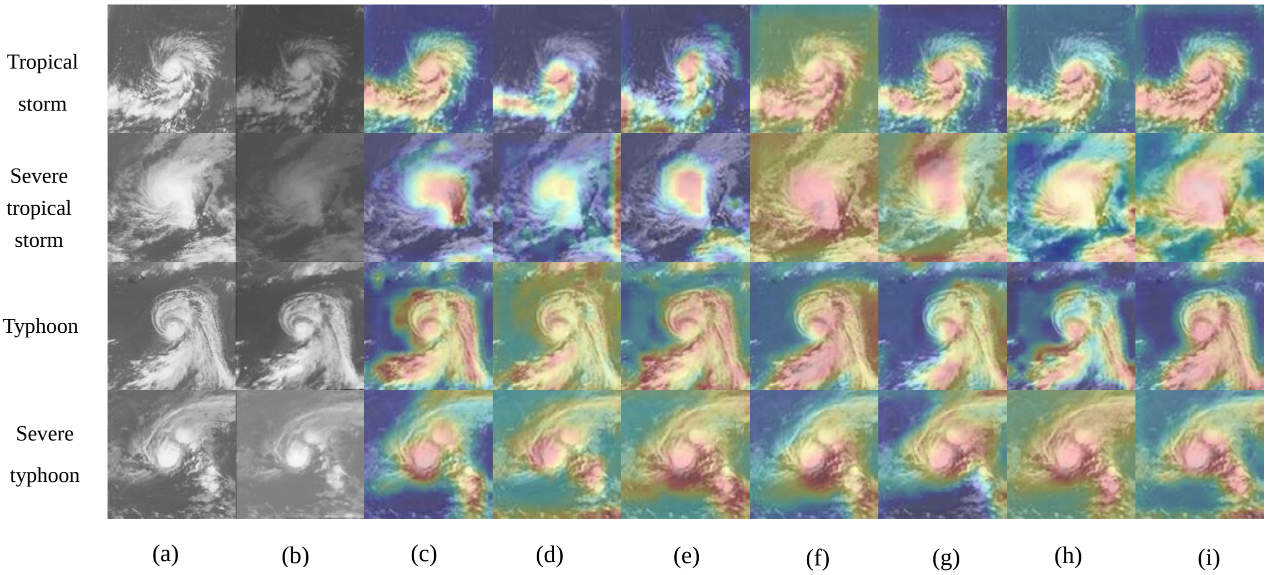


Figure 12. Comparison of the Grad-CAM heatmaps under different registration + fusion methods under four categories: (a) HMW cloud images, (b) FY cloud images, (c) Coarse-to-fine registration +DAE+CMFI, (d) DAE+CMFI, (e) VoxelMorph + DAE+CMFI, (f) SIFT +DAE+CMFI, (g) Coarse-to-fine registration + IFCNN, (h) VoxelMorph + IFCNN, and (i) SIFT + IFCNN.

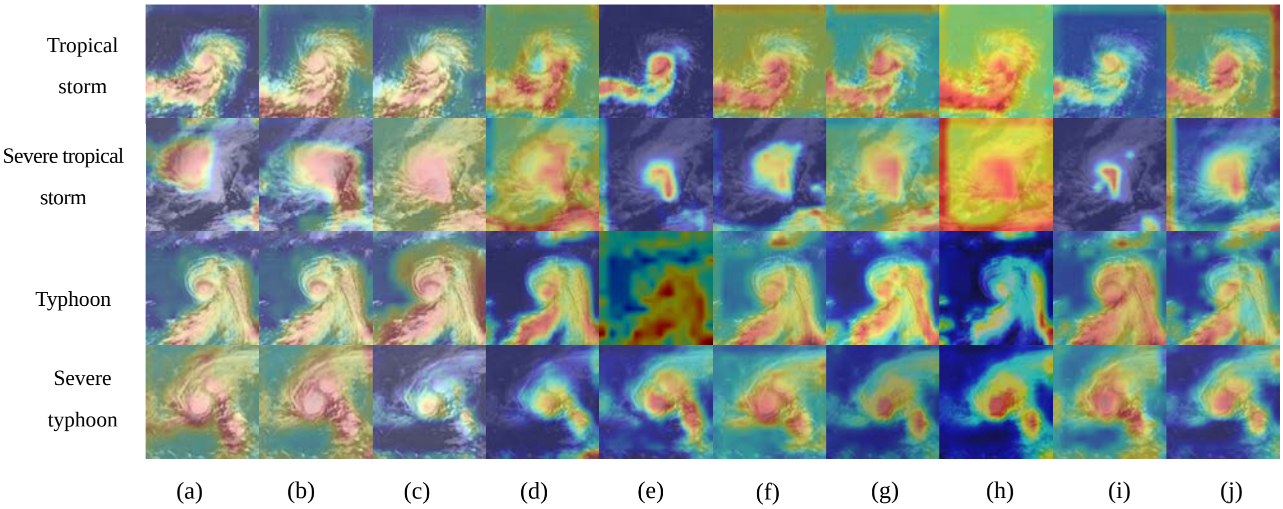


Figure 13. Comparison of the Grad-CAM heatmaps of different methods: (a) Coarse-to-fine registration + DenseFuse, (b) VoxelMorph + DenseFuse, (c) SIFT + DenseFuse, (d) Bi-JROS + IFCNN, (e) Bi-JROS + DenseFuse, (f) Bi-JROS + SPFusion, (g) Bi-JROS + DAE+CMFI, (h) SIFT + SPFusion, (i) VoxelMorph + SPFusion, and (j) Coarse-to-fine registration + SPFusion.

4. Conclusions

This paper introduces a novel one-stage coarse-to-fine registration and fusion network designed to enhance typhoon cloud image classification using multimodal satellite imagery. The proposed framework addresses the challenge of image misalignment by integrating a one-stage coarse-to-fine registration module with a cross-modal fusion method. The registration module effectively aligns multimodal feature data extracted by a two-branch convolutional neural network, achieving both image distribution alignment and geometric structure alignment. This alignment process mitigates radiometric and spatial discrepancies between images from different satellites. The fusion module applies a bidirectional attention mechanism to reduce redundant features while enhancing feature utilization across modalities. By incorporating a residual structure, the fusion module optimizes the use of low-level, high-level, intra-modal, and intermodal features, leading to improved classification accuracy.

Ablation experiments confirm the effectiveness of combining the registration and fusion modules. Specifically, this joint approach improved classification accuracy to 84.34% on the FY dataset and 82.88% on the HMW dataset. Comparative experiments with other registration and fusion methods demonstrate that the proposed one-stage coarse-to-fine registration, paired with the DAE+CMFI fusion method, achieves superior classification performance. These results highlight the importance of addressing image distribution differences during the registration process and underscore the value of reducing information redundancy during fusion. Enhancing feature interaction across modalities significantly improves classification accuracy for multimodal typhoon cloud images.

In future work, we plan to expand the dataset by incorporating additional modalities, particularly image and text data, which present a larger modality gap. Additionally, we aim to optimize the neural network architecture to extract more discriminative and robust features, thereby enhancing the accuracy of typhoon cloud image classification. Moreover, we will investigate the integration of more robust denoising methods to improve noise handling and address severe data misalignment issues in real-world environments. Furthermore, we plan to explore the application of CoReFuNet to different weather phenomena, such as hurricanes and tornadoes, to assess its adaptability to various meteorological conditions. In addition, we aim to investigate its effectiveness across different satellite sensing platforms, to ensure robustness under varying spatial and temporal resolutions.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

This study was supported by the NSF China [grant numbers 41671431]; Shanghai Municipal Science and Technology Commission Local Colleges Capacity Building Project [grant numbers 19050502100]; and Guangzhou Meteorological Satellite Ground Station Project [grant number D-8006-23-0157].

Conflict of interest

The authors declare there is no conflict of interest.

References

1. X. Sun, Y. Tian, W. Lu, P. Wang, R. Niu, H. Yu, et al., From single-to multi-modal remote sensing imagery interpretation: A survey and taxonomy, *Sci. China Inf. Sci.*, **66** (2023), 140301. https://doi.org/10.1007/s11432-022-3588-0
2. L. Li, H. Ling, M. Ding, H. Cao, H. Hu, A deep learning semantic template matching framework for remote sensing image registration, *ISPRS*, **181** (2021), 205–217. https://doi.org/10.1016/j.isprsjprs.2017.12.012
3. R. Hang, Z. Li, P. Ghamisi, D. Hong, G. Xia, Q. Liu, Classification of hyperspectral and LiDAR data using coupled CNNs, *IEEE Trans. Geosci. Remote Sens.*, **58** (2020), 4939–4950. https://doi.org/10.1109/TGRS.2020.2969024
4. S. Morchhale, V. P. Pauca, R. J. Plemmons, T. C. Torgersen, Classification of pixel-level fused hyperspectral and lidar data using deep convolutional neural networks, in *2016 8th Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS)*, (2016), 1–5. https://doi.org/10.1109/WHISPERS.2016.8071715
5. S. Deldari, H. Xue, A. Saeed, D. V. Smith, F. D. Salim, Cocoa: Cross modality contrastive learning for sensor data, *Proc. ACM Interact. Mobile Wearable Ubiquitous Technol.*, **6** (2022), 1–28. https://doi.org/10.1145/3550316
6. D. Hong, L. Gao, N. Yokoya, J. Yao, J. Chanussot, Q. Du, et al., More diverse means better: Multimodal deep learning meets remote-sensing imagery classification, *IEEE Trans. Geosci. Remote Sens.*, **59** (2020), 4340–4354. https://doi.org/10.1109/TGRS.2020.3016820
7. H. Li, X. Wu, DenseFuse: A fusion approach to infrared and visible images, *IEEE Trans. Image Process.*, **28** (2018), 2614–2623. https://doi.org/10.1109/TIP.2018.2887342
8. Y. Zhang, Y. Liu, P. Sun, H. Yan, X. Zhao, X. Zhang, IFCNN: A general image fusion framework based on convolutional neural network, *Inf. Fusion*, **54** (2020), 99–118. https://doi.org/10.1016/J.INFFUS.2019.07.011
9. H. Xu, J. Ma, J. Jiang, X. Guo, H. Ling, U2Fusion: A unified unsupervised image fusion network, *IEEE Trans. Pattern Anal. Mach. Intell.*, **44** (2020), 502–518. https://doi.org/ 10.1109/TPAMI.2020.3012548
10. N. Li, Y. Li, J. Jiao, Multimodal remote sensing image registration based on adaptive multi-scale PIIFD, *Multimed. Tools Appl.*, **83** (2024), 1–13. https://doi.org/10.1007/s11042-024-18756-1
11. H. Xie, J. Qiu, Y. Dai, Y. Yang, C. Cheng, Y. Zhang, SA-DNet: A on-demand semantic object registration network adapting to non-rigid deformation, preprint, arXiv:2210.09900. https://doi.org/10.48550/arXiv.2210.09900
12. R. Feng, H. Shen, J. Bai, X. Li, Advances and opportunities in remote sensing image geometric registration: A systematic review of state-of-the-art approaches and future research directions, *IEEE Trans. Geosci. Remote Sens.*, **9** (2021), 120–142. https://doi.org/ 10.1109/MGRS.2021.3081763
13. W. M. Wells III, P. Viola, H. Atsumi, S. Nakajima, R. Kikini., Multi-modal volume registration by maximization of mutual information, *Med. Image Anal.*, **1** (1996), 35–51. https://doi.org/10.1016/s1361-8415(01)80004-9
14. A. Goshtasby, G. C. Stockman, C. V. Page, A region-based approach to digital image registration with subpixel accuracy, *IEEE Trans. Geosci. Remote Sens.*, **24** (1986), 390–399. https://doi.org/ 10.1109/TGRS.1986.289597
15. X. He, C. Meile, S. M. Bhandarkar, Multimodal registration of FISH and nanoSIMS images using convolutional neural network models, preprint, arXiv:2201.05545. https://doi.org/10.48550/arXiv.2201.05545

16. D. G. Lowe, Object recognition from local scale-invariant features, in *Proceedings of the Seventh IEEE International Conference on Computer Vision*, **2** (1999), 1150–1157. https://doi.org/10.1109/ICCV.1999.790410
17. J. Li, Q. Hu, M. Ai, RIFT: Multi-modal image matching based on radiation-variation insensitive feature transform, *IEEE Trans. Image Process.*, **29** (2019), 3296–3310. https://doi.org/10.1109/TIP.2019.2959244
18. L. Tang, Y. Deng, Y. Ma, J. Huang, J. Ma, SuperFusion: A versatile image registration and fusion network with semantic awareness, *IEEE/CAA J. Autom. Sin.*, **9** (2022), 2121–2137. https://doi.org/10.1109/JAS.2022.106082
19. B. K. P. Horn, B. G. Schunck, Determining optical flow, *Artif. Intell.* **17** (1981), 185–203. https://doi.org/10.1016/0004-3702(81)90024-2
20. J. Xiong, Y. Luo, G. Tang, An improved optical flow method for image registration with large-scale movements, *Acta Autom. Sin.*, **34** (2008), 760–764. https://doi.org/10.3724/SP.J.1004.2008.00760
21. H. Li, J. Zhao, J. Li, Z. Yu, G. Lu, Feature dynamic alignment and refinement for infrared-visible image fusion: Translation robust fusion, *Inf. Fusion*, **95** (2023), 26–41. https://doi.org/10.1016/j.inffus.2023.02.011
22. S. Tang, P. Miao, X. Gao, Y. Zhong, D. Zhu, H. Wen, et al., Point cloud-based registration and image fusion between cardiac SPECT MPI and CTA, preprint, arXiv:2402.06841. https://doi.org/10.48550/arXiv.2402.06841
23. Z. Xu, J. Yan, J. Luo, X. Li, J. Jagadeesan, Unsupervised multimodal image registration with adaptative gradient guidance, in *ICASSP 2021–2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, (2021), 1225–1229. https://doi.org/10.1109/ICASSP39728.2021.9414320
24. H. Xu, J. Ma, J. Yuan, Z. Le, W. Liu, RFNet: Unsupervised network for mutually reinforcing multi-modal image registration and fusion, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2022), 19647–19656. https://doi.org/10.1109/CVPR52688.2022.01906
25. H. Xu, J. Yuan, J. Ma, Murf: Mutually reinforcing multi-modal image registration and fusion, *IEEE Trans. Pattern Anal. Mach. Intell.*, **45** (2023), 12148–12166. https://doi.org/10.1109/TPAMI.2023.3283682
26. D. Wang, J. Liu, X. Fan, R. Liu, Unsupervised misaligned infrared and visible image fusion via cross-modality image generation and registration, preprint, arXiv:2205.11876. https://doi.org/10.48550/arXiv.2205.11876
27. D. Wang, J. Liu, L. Ma, R. Liu, X. Fan, Improving misaligned multi-modality image fusion with one-stage progressive dense registration, *IEEE Trans. Circuits Syst. Video Technol.*, **34** (2024). https:// doi.org/10.1109/TCSVT.2024.3412743
28. Z. Zhang, H. Li, T. Xu, X. Wu, J. Kittler, BusReF: Infrared-visible images registration and fusion focus on reconstructible area using one set of features, preprint, arXiv:2401.00285. https://doi.org/10.48550/arXiv.2401.00285
29. L. Z. Li, L. Han, M. Ding, H. Cao, Multimodal image fusion framework for end-to-end remote sensing image registration, *IEEE Trans. Geosci. Remote Sens.*, **61**, (2023), 1–14. https://doi.org/10.1109/TGRS.2023.3247642
30. S. Mai, Y. Zeng, H. Hu, Multimodal information bottleneck: Learning minimal sufficient unimodal and multimodal representations, *IEEE Trans. Multimedia*, **25** (2022), 4121–4134. https://doi.org/10.1109/TMM.2022.3171679

31. Q. Wang, Y. Chi, T. Shen, J. Song, Z. Zhang, Y. Zhu, Improving RGB-infrared object detection by reducing cross-modality redundancy, *Remote. Sens.*, **14** (2020). https://doi.org/10.3390/rs14092020
32. S. Cui, J. Cao, X. Cong, J. Sheng, Q. Li, T. Liu, et al., Enhancing multimodal entity and relation extraction with variational information bottleneck, *IEEE/ACM Trans. Audio Speech Lang. Process.*, **32** (2024), 1274–1285. https://doi.org/10.1109/TASLP.2023.3345146
33. H. Chen, Y. Li, Three-stream attention-aware network for RGB-D salient object detection, *IEEE Trans. Image Process.*, **28** (2019), 2825–2835. https://doi.org/10.1109/TIP.2019.2891104
34. X. Sun, L. Zhang, H. Yang, T. Wu, Y. Cen, Y. Guo, Enhancement of spectral resolution for remotely sensed multispectral image, *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.*, **8** (2014), 2198–2211. https://doi.org/10.1109/JSTARS.2014.2356512
35. S. Malec, D. Rogge, U. Heiden, A. Sanchez-Azofeifa, M. Bachmann, M. Wegmann, Capability of spaceborne hyperspectral EnMAP mission for mapping fractional cover for soil erosion modeling, *Remote Sens.*, **7** (2015), 11776–11800. https://doi.org/10.3390/rs70911776
36. N. Yokoya, T. Yairi, A. Iwasaki, Coupled nonnegative matrix factorization unmixing for hyperspectral and multispectral data fusion, *IEEE Trans. Geosci. Remote Sens.*, **50** (2020), 528–537. https://doi.org/10.1109/TGRS.2011.2161320
37. L. Mou, X. Zhu, RiFCN: Recurrent network in fully convolutional network for semantic segmentation of high resolution remote sensing images, preprint, arXiv:1805.02091. https://doi.org/10.48550/arXiv.1805.02091
38. H. Ma, X. Yang, R. Fan, W. Han, K. He, L. Wang, Refined water-body types mapping using a water-scene enhancement deep models by fusing optical and SAR images, *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, **17** (2024), 17430–17441. https://doi.org/10.1109/JSTARS.2024.3459916
39. D. Hong, N. Yokoya, J. Chanussot, X. Zhu, CoSpace: Common subspace learning from hyperspectral-multispectral correspondences, *IEEE Trans. Geosci. Remote Sens.*, **57** (2019), 4349–4359. https://doi.org/10.1109/TGRS.2018.2890705
40. D. Hong, N. Yokoya, N. Ge, J. Chanussot, X. Zhu, Learnable manifold alignment (LeMA): A semi-supervised cross-modality learning framework for land cover and land use classification, *ISPRS J. Photogramm. Remote Sens.*, **147** (2019), 193–205. https://doi.org/10.1016/j.isprsjprs.2018.10.006
41. G. Li, Z. Liu, M. Chen, Z. Bai, W. Lin, H. Ling, Hierarchical alternate interaction network for RGB-D salient object detection, *IEEE Trans. Image Process.*, **30** (2021), 3528–3542. https://doi.org/10.1109/TIP.2021.3062689
42. M. Chen, L. Xing, Y. Wang, Y. Zhang, Enhanced multimodal representation learning with cross-modal kd, in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2023), 11766–11775.
43. D. Hong, J. Yao, D. Meng, Z. Xu, J. Chanussot, Multimodal GANs: Toward crossmodal hyperspectral-multispectral image segmentation, *IEEE Trans. Geosci. Remote Sens.*, **59** (2020), 5103–5113. https://doi.org/10.1109/TGRS.2020.3020823
44. D. Hong, N. Yokoya, G. Xia, J. Chanussot, X. Zhu, X-ModalNet: A semi-supervised deep cross-modal network for classification of remote sensing data, *ISPRS J. Photogramm. Remote Sens.*, **167** (2020), 12–23. https://doi.org/10.1016/j.isprsjprs.2020.06.014
45. J. Huang, X. Huang, J. Yang, Residual enhanced multi-hypergraph neural network, in *2021 IEEE International Conference on Image Processing (ICIP)*, (2021), 3657–3661. https://doi.org/10.1109/ICIP42928.2021.9506153

46. N. Xu, W. Mao, A residual merged neutral network for multimodal sentiment analysis, in *2017 IEEE 2nd International Conference on Big Data Analysis (ICBDA)*, (2017), 6–10. https://doi.org/10.1109/ICBDA.2017.8078794
47. K. He, Z. Zhang, Y. Dong, D. Cai, Y. Lu, W. Han, Improving geological remote sensing interpretation via a contextually enhanced multiscale feature fusion network, *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, **17** (2024), 6158–6173. https://doi.org/10.1109/JSTARS.2024.3374818
48. A. M. Saxe, Y. Bansal, J. Dapello, M. Advani, A. Kolchinsky, B. D. Tracey, et al., On the information bottleneck theory of deep learning, *J. Stat. Mech.: Theory Exp.*, **2019** (2019), 124020. https://doi.org/10.1088/1742-5468/ab3985
49. B. Chen, B. Chen, H. Lin, R. L. Elsberry, Estimating tropical cyclone intensity by satellite imagery utilizing convolutional neural networks, *Weather Forecast.*, **34** (2019), 447–465. https://doi.org/10.1007/s13351-024-3186-y
50. J. Lee, J. Im, D. H. Cha, H. Park, S. Sim, Tropical cyclone intensity estimation using multi-dimensional convolutional neural networks from geostationary satellite data, *Remote Sens.*, **12** (2019), 108. https://doi.org/10.3390/rs12010108
51. R. Zhang, Q. Liu, R. Hang, Tropical cyclone intensity estimation using two-branch convolutional neural network from infrared and water vapor images, *IEEE Trans. Geosci. Remote Sens.*, **58** (2019), 586–597. https://doi.org/10.1109/TGRS.2019.2938204
52. W. Jiang, G. Hu, T. Wu, L. Liu, B. Kim, Y. Xiao, et al., DMANet_KF: Tropical cyclone intensity estimation based on deep learning and Kalman filter from multi-spectral infrared images, *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, **16** (2023), 4469–4483. https://doi.org/10.1109/JSTARS.2023.3273232
53. M. Brown, S. Süsstrunk, Multi-spectral SIFT for scene category recognition, in *CVPR 2011*, (2011), 177–184. https://doi.org/10.1109/CVPR.2011.5995637
54. L. Tang, H. Zhang, H. Xu, J. Ma, Rethinking the necessity of image fusion in high-level vision tasks: A practical infrared and visible image fusion network based on progressive semantic injection and scene fidelity, *Inf. Fusion*, **99** (2023), 101870. https://doi.org/10.1016/j.inffus.2023.101870
55. G. Balakrishnan, A. Zhao, M. R. Sabuncu, J. Guttag, A. V. Dalca, Voxelmorph: A learning framework for deformable medical image registration, *IEEE Trans. Med. Imaging*, **38** (2019), 1788–1800. https://doi.org/10.1109/TMI.2019.2897538
56. X. Fan, X. Wang, J. Gao, J. Wang, Z. Luo, R. Liu, Bi-level learning of task-specific decoders for joint registration and one-shot medical image segmentation, in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2024), 11726–11735. https://doi.org/10.1109/CVPR52733.2024.01114
57. R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-cam: Visual explanations from deep networks via gradient-based localization, in *Proceedings of the IEEE International Conference on Computer Vision*, (2017), 618–626. https://doi.org/10.1109/ICCV.2017.74
58. V. F. Dvorak, Tropical cyclone intensity analysis and forecasting from satellite imagery, *Mon. Weather Rev.*, **103** (1975), 420–430. https://doi.org/10.1175/1520-0493(1975)103%3C0420:TCIAAF%3E2.0.CO;2