Unsupervised Multimodal Remote Sensing Image Registration via Domain Adaptation

Lukui Shi, Ruiyun Zhao, Bin Pan, *Member, IEEE*, Zhengxia Zou,
and Zhenwei Shi, *Senior Member, IEEE*

***Abstract*— Registration of multimodal remote sensing images with geometric distortions is one of the fundamental applications, but it remains difficult since multimodal remote sensing images have significant differences in both radiometric and geometric features. One of the challenges is the disregarding of modality-specific information, which hinders the model from focusing on the content information of structure and texture due to differences in radiometric features. In this article, an unsupervised content-focused hierarchical alignment network (CHA-Net) is proposed, which is constructed based on the theory of domain adaptation. The kernel idea of CHA-Net is to weaken the style differences among different modal images and achieve nonrigid multimodal remote sensing image registration. CHA-Net is a hierarchical refinement model, where different scales of features are aligned, respectively, by utilizing the field calibration module (FCM) and gradually generating the registration field. To be specific, CHA-Net consists of two structures: the Siamese feature decoupling (SFD) structure and the hierarchical refinement alignment (HRA) structure. The SFD aims at reducing the style differences caused by cross-modal differences and developing a shared-weight Siamese network to map images to content feature space. The HRA enhances the ability of the network by capturing global distortions based on the transformer model. Experiments on public datasets indicate that compared with other methods, CHA-Net performs better when geometric and radiometric distortions appear.**

***Index Terms*— Domain adaptation, hierarchical alignment, multimodal image registration, remote sensing images.**

I. Introduction

MULTIMODAL image registration refers to the process of spatially aligning corresponding parts of remote sensing images obtained from different sensors, which exhibit geometric and radiometric differences [1]. With the rapid development of remote sensing technology, various sensors are becoming increasingly capable of observing the Earth, and many remote sensing applications, such as change detection, image fusion, and image stitching, require collaborative analysis of multiscene remote sensing images, which cannot be achieved without remote sensing image registration [2], [3]. Multimodal image registration has always been the core problem of remote sensing image registration. Due to different imaging mechanisms, multimodal images often exhibit significant and independent features, making it difficult to measure the degree of geometric alignment, thus making it challenging to achieve registration between different modalities of images.

Over the past decade, deep learning-based methods have been proposed for image registration tasks [4], [5], [6], [7], [8]. These methods can extract distributions from data and learn deep-level information. They can be categorized into supervised and unsupervised learning approaches. Supervised learning methods require a large number of labels for training, which can be expensive and difficult to obtain for registration tasks [9], [10]. On the other hand, unsupervised learning methods, which possess the characteristic of being label-free, have been extensively studied in the registration field.

Remote sensing employs diverse sensors on satellites or airplanes to capture images of the Earth, including various modalities of imagery. These images are utilized in ecosystem evolution monitoring [11], urban development, and automated map creation [12]. However, factors like spatial resolution disparities and viewpoint differences necessitate image registration [13]. Solely relying on rigid registration [14], which focuses on global transformations, is insufficient to achieve accurate image alignment. Additional correction is required at the local level, calling for nonrigid registration [4], [15] that demands a more refined approach. This study aims to investigate and develop nonrigid registration algorithms.

The field of multimodal image registration using deep learning techniques encompasses three commonly employed strategies [1]. First, constructing deep and advanced image feature representations to acquire recurrent feature points and/or advanced descriptors for feature matching [16], [17], [18], [19], [20], [21]. To establish the correspondence between features in multimodal image pairs, HOPC descriptor [22] was proposed. It represents the characteristics of local regions by calculating the principal component direction and strength, based on the histogram of oriented phase congruency.

Manuscript received 10 July 2023; revised 7 October 2023; accepted 12 November 2023. Date of publication 16 November 2023; date of current version 29 November 2023. This work was supported in part by the National Key Research and Development Program of China under Grant 2022ZD0160401 and Grant 2022YFA1003803, in part by the Scientific and Technological Research Project of Hebei Province Universities and Colleges under Grant ZD2021311, in part by the National Natural Science Foundation of China under Grant 62001251, and in part by the Beijing-Tianjin-Hebei Basic Research Cooperation Project under Grant F2021203109. *(Corresponding author: Bin Pan.)*

Lukui Shi and Ruiyun Zhao are with the School of Artificial Intelligence, Hebei University of Technology, Tianjin 300401, China, and also with the Hebei Province Key Laboratory of Big Data Calculation, Tianjin 300401, China (e-mail: shilukui@scse.hebut.edu.cn; ruiyunzhao@outlook.com).

Bin Pan is with the School of Statistics and Data Science, KLMDASR, LEBPS, and LPMC, Nankai University, Tianjin 300071, China (e-mail: panbin@nankai.edu.cn).

Zhengxia Zou is with the Department of Guidance, Navigation and Control, School of Astronautics, Beihang University, Beijing 100191, China (e-mail: zhengxiazou@buaa.edu.cn).

Zhenwei Shi is with the Image Processing Center, School of Astronautics, Beihang University, Beijing 100191, China (e-mail: shizhenwei@buaa.edu.cn).

Digital Object Identifier 10.1109/TGRS.2023.3333889

Wang et al. [23] partitioned images from various modalities into multiple patches to learn the mapping between these paired patches and their corresponding matching labels, thereby generating a mapping function for interfeature relationships. Second, establishing a mapping relationship between imagery and either homography matrix parameters or displacement field models through end-to-end networks [24], [25], [26]. Papadomanolaki et al. [15] proposed a multistep nonrigid registration approach based on convolutional neural networks (CNNs). In one registration process, iterations are performed repeatedly through the same U-net to capture the geometric distortions present in the remote sensing images. NBR-Net [4] is proposed for remote sensing images to predict the flow field caused by distortion and viewpoint difference. It enhances registration reversibility and geometric consistency by applying a cyclic registration strategy, which registers the registered image back to the original image. With the rapid development of transformer models in deep learning, Transmorph [27] is proposed, which leverages transformer [28] as a spatial encoder. It effectively captures the geometric distortions present between image pairs and thoroughly demonstrates its effectiveness in registration models through extensive experiments. Third, employing networks such as GAN [29] to map images into a unified modality, thereby suppressing feature disparities arising from imaging mechanisms or other factors [30], [31]. Ma et al. [32] enhance the structural coherence between image pairs by utilizing CNN to extract deep features, thereby strengthening the matching correlation between features. They demonstrate the effectiveness of capturing spatial structural characteristics for registration from multimodal images. CGRP [33] constructed an image translation network and a registration network. They utilized the image translation network to transform images of different modalities into the same modality, treating it as a monomodal registration problem by leveraging geometric preservation. NeMAR [34] was proposed, where an image translation network and a registration network are jointly trained in an alternating manner. This allows the registration network to focus on capturing geometric distortions in multimodal images, enabling end-to-end registration using only the registration network.

However, previous registration methods for remote sensing images based on deep learning have overlooked the core issue of inconsistent distributions caused by imaging style differences in multimodal remote sensing images. Therefore, when facing various complex remote sensing scenarios, it is prone to produce distorted images that are difficult to maintain robustness.

Recently, domain adaptation methods [35], [36] have made significant progress in overcoming the problem of inconsistent distributions between source and target domains, which is similar to the problem of different imaging styles in multimodal image registration. Inspired by domain adaptation ideas, we aim to create an image registration method for multimodal images based on domain adaptation. However, domain adaptation can only improve distribution alignment issues and cannot resolve spatial geometric differences between multimodal image pairs.

In this article, we present an unsupervised Content-focused Hierarchical Alignment network (CHA-Net) for multimodal image registration, which achieves alignment of both image distributions and image geometric structures. The proposed algorithm comprises two main components: the Siamese feature decoupling (SFD) structure and the hierarchical refinement alignment (HRA) structure. The SFD effectively reduces radiation distortion caused by style differences in multimodal images. It achieves this by mapping the images to a content feature space, causing the network to focus on the content information in the images. The HRA structure enhances the ability of the network to capture global distortions utilizing the transformer model, while still retaining its capability to capture local distortions and facilitating alignment at different scales of features.

The contributions in this article can be summarized as follows.

1) We design an end-to-end nonrigid multimodal remote sensing image registration network (CHA-Net) and validate its effectiveness on multiple datasets.
2) We propose the SFD structure to effectively weaken the style differences among multimodal images and achieve distribution alignment between them.
3) We propose the HRA structure to enhance the global modeling ability of the algorithm and leverage the field calibration module (FCM) to achieve nonrigid registration in a hierarchical refinement manner.

II. Methodology

In this section, we will present the framework of the algorithm, as shown in Fig. 1. We will begin by defining the task of multimodal remote sensing image registration, followed by an overview of the overall architecture and strategy of the CHA-Net. We will then describe the structures of the main modules of the model, and conclude with an explanation of the optimization method employed by our model.

A. Problem Definition

Image registration refers to the establishment of spatial correspondences between two images by optimizing an energy function to maintain geometric alignment between corresponding points in the fixed image and the moving image. This process can be mathematically represented by the following equation:

$$\min E(I_m, I_f, \phi) = \mathrm{Sim}(I_m \circ \phi, I_f) + \lambda \nabla(\phi). \tag{1}$$

In this equation, I_m and I_f denote the moving and fixed images, respectively, where the former is the image to be registered that presents geometrical distortions and the latter is the reference image. ϕ signifies the registration field that aligns the moving image with the fixed image, $\nabla(\phi)$ refers to the regularization term used to ensure that the deformation field does not create distortions. Moreover, λ denotes the regularization coefficient, E represents the idealized similarity function, and Sim describes the alignment level between the transformed moving image and the fixed image. The objective of image registration is to learn a nonlinear transformation ϕ

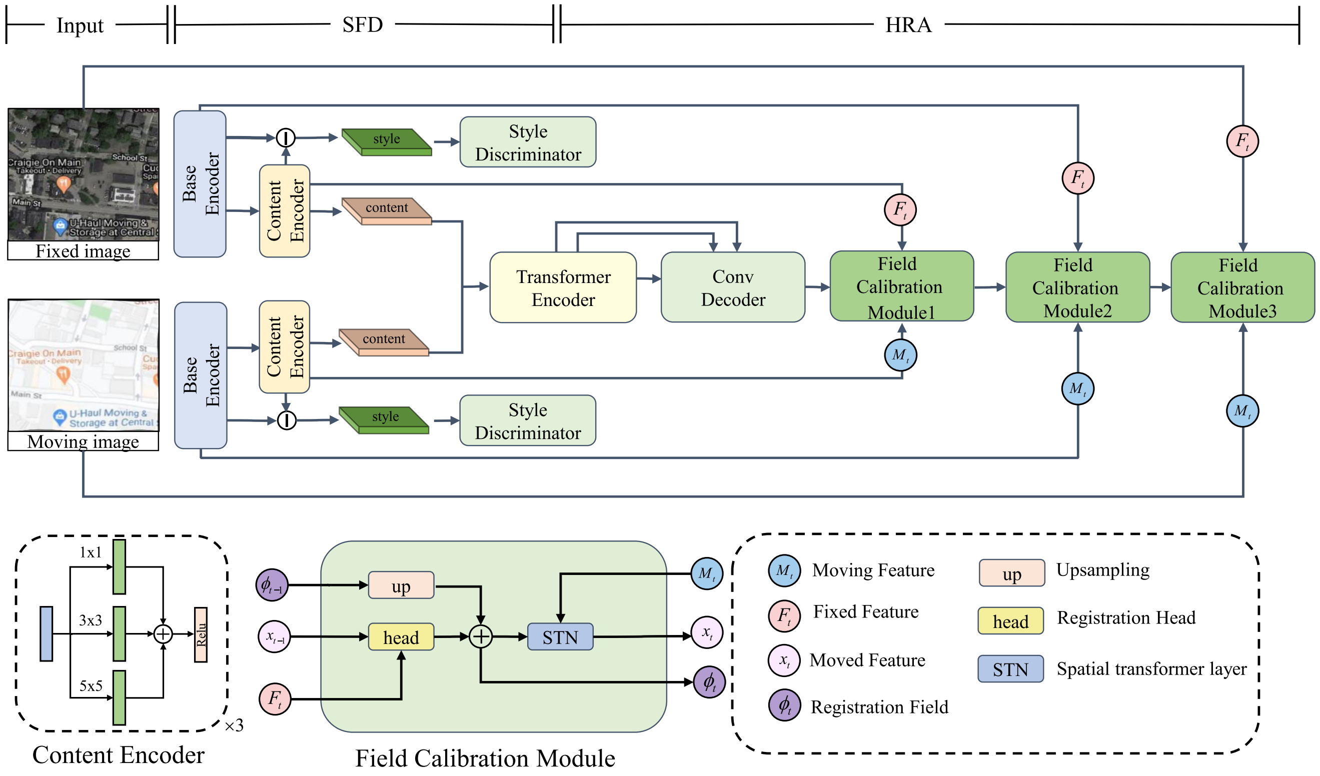


Fig. 1. Proposed unsupervised content-focused hierarchical alignment network, CHA-Net, for misaligned multimodal image pairs consists of two main components: the SFD structure and the HRA structure. This method takes misaligned multimodal image pairs as input and passes them sequentially through the SFD and HRA structures to obtain the final alignment results.

by optimizing the similarity function, which aligns two images originating from different modalities

$$\hat{\phi} = \arg\min E(I_m, I_f, \phi). \qquad (2)$$

End-to-end methods aim to directly predict the registration field ϕ and then perform spatial geometric alignment based on the deformation field ϕ using the spatial transformer networks (STNs) [37]. As the STN is differentiable, it can leverage the backpropagation mechanism for end-to-end learning.

B. Framework Overview

The overall architectural design of the model can be visualized in Fig. 1, The CHA-Net is composed of two integral parts: the SFD structure and the HRA structure. Together, they form a nonrigid multimodal remote sensing image registration network. The SFD enables the mapping of images to a content feature space, thereby reducing the distance between different modalities and enhancing the content information of the image while eliminating feature disturbances caused by radiation distortion. The HRA employs the transformer to process decoupled features and estimate the global registration field, which is then corrected iteratively through the FCM. The structure retains local modeling capability while enhancing the model's global distortion capture ability through a transformer, thereby enabling hierarchical alignment at different scales of features

C. SFD Structure

Compared to registering single-modality images, multimodality image pairs exhibit significant differences in both imaging bands and imaging mechanisms, making it challenging to find a nonrigid transformation that aligns two images from different modalities. Although there are differences between the images from different modalities, there are still many commonalities between them, such as the texture and spatial structure of the depicted objects in the images. Therefore, in this proposed approach, we assume that an image is composed of two distinct parts: content information and style information. The style information reflects the modality-specific characteristics such as radiance and color features of the image, whereas the content information primarily stands for the structural texture features of the image. In order to accurately identify and rectify geometric distortions, when the model is matching geometric distortions, it is critical to focus on the content information rather than the style information. Style information tends to act as noise, leading the registration to proceed in the wrong direction and making the registration task arduous.

Our core idea is to enable the model to focus more on the content information between different modality images. By reducing the feature distribution distance between different modalities, we can reduce the modality-specific style information. To achieve this, we have designed the SFD to reduce the influence of different modalities and mitigate feature disturbances caused by radiation distortion.

The SFD consists of a pair of dual-branch networks with shared weights, with each branch composed of the base encoder (BE), the content encoder (CE), and the style discriminator (SD). The dual-branch structure consists of three modules in each branch, which share weights with each other, facilitating the extraction of shared information from the input feature. The BE is composed of CNNs, serving as a basic feature extractor. The CE consists of three multikernel convolutional layers, with each layer containing multiple convolutional kernels of different sizes. This design enhances the diversity of gradients while maintaining the same size of input and content features, as illustrated in Fig. 1. To enable the SFD to focus on content extraction, it is necessary to first enable it to extract style information. Therefore, we have included the SD, which trains the network to identify the source of the style information from which modality, thereby endowing the network with the capability of capturing style information.

Taking the single-branch SFD as an example, the process is described as follows. First, we select one image from the multimodality image pair as the input and utilize the BE to extract the basic features $F \in R^{(C \times H \times W)}$, where C, H, W denote the number of channels, height, and width, respectively. By CE we can decompose F into the content feature map $C \in R^{(C \times H \times W)}$ and the style feature map $S \in R^{(C \times H \times W)}$ such that $F = C + S$, This process can be mathematically expressed as follows:

$$C = \mathcal{M}(F), S = F - C. \tag{3}$$

In this formula, $\mathcal{M}$ represents the CE responsible for obtaining the style features and inputting them into SD for supervision training. Through the process of backpropagation, the model can learn style information focused on specific features of multimodality. To ensure that the acquired content features remain focused on structural textures rather than stylistic features such as radiation, it is crucial to maintain the independence between content and style features. This is achieved by minimizing the mutual information (MI) loss between content and style features, which represents the extent of their unrelatedness. The MI loss is expressed as follows:

$$L_{\mathrm{MI}} = \int_{A \times B} \ln \frac{\mathrm{d}\mathbb{P}_{AB}}{\mathrm{d}\mathbb{P}_A \otimes \mathrm{d}\mathbb{P}_B} \mathrm{d}\mathbb{P}_{AB}. \tag{4}$$

Here, P_{AB} represents the joint probability distribution of features A and B, while P_A and P_B represent marginal probability distributions. The SFD employs domain adaptation to minimize feature distances between different modalities, ensuring that the same features have more similar distributions after being extracted by the Siamese network structure. When the boundaries between different modal distributions become blurred, the implicit style distributions in multimodal images will be weakened, enabling the model to focus on extracting content information to improve registration performance.

D. HRA Structure

As SFD weakens the influence across modalities, another core component of registration is how to obtain the registration field ϕ. Our aim is to learn a nonrigid transformation that uses the fixed image as the basis to generate the registration field ϕ based on the differences between the moving image and the fixed image. By combining with the STN for resampling the moving image, the registered image can be obtained.

Previous image registration models often utilized multiscale fusion or multistep iterations in a coarse-to-fine manner to enhance the accuracy of registration. However, these methods lacked attention to the characteristic features of different hierarchical levels, neglecting the importance of geometric alignment of the feature maps generated during the registration process. To better capture the nonrigid distortions present in remote sensing images, we have designed the HRA structure. HRA aims to focus on the characteristics of different hierarchical features and perform varying degrees of alignment on them through STN.

The HRA comprises a transformer-based field estimator and multiple FCM, wherein the transformer-based field estimator utilizes Swin-transformer as the encoder and a CNN as the decoder, integrated with a U-net architecture that facilitates the capture of distortion relationships at the global scale, which is crucial for generating the first-level registration field. The FCM primarily consists of a Registration Head and the STN as shown in Fig. 1, which is based on the deep-level features x_{t-1} and the previous layer's features $\{M_t, F_t\}$, corrects the registration field ϕ_{t-1} generated from the previous layer. This enables multiple corrections of the registration field using FCM and multiple resampling of features, thereby achieving hierarchical alignment of features at various scales.

To be more specific, HRA takes the content feature obtained from the previous layer and inputs it into the transformer-based field estimator. At low resolutions, it captures the distortions of the moving image and fixed image on a global scope, and combines with FCM to estimate the first-level registration field ϕ_1. Based on the results of the previous layer's registration field, it resamples the features of the moving image at this stage to achieve spatial alignment of the features. Despite this, the registration field at this stage still contains many errors. By utilizing FCM to repeatedly correct the erroneous distortions existing in ϕ_t, and resampling the initial features of higher-resolution images, it is possible to capture nongrid geometric distortions and generate a more accurately aligned registration field ϕ at the pixel level. The process of FCM can be represented by the following equation:

$$\phi_t = \mathrm{CAL}(F_t, M_t, \phi_{t-1}) + \phi_{t-1} \tag{5}$$

where $\{F_t, M_t\}$, respectively, represent the features at level t of the network, CAL represents the correction made to the existing registration field ϕ based on features $\{F_t, M_t\}$. Through multistep calibration, we can refine the registration field and obtain the desired ϕ when $t = T$. Furthermore, to further align the obtained features, the hierarchical alignment loss is designed to ensure that the resampled features remain aligned. It is represented as follows:

$$L_{\mathrm{ha}} = \sum_t \|\phi_t \circ F_t - M_t\|_1. \tag{6}$$

Notably, despite the significant parameter and computational resource requirements of the transformer, the HRA extracts low-resolution, multichannel features, enabling effective capture of global image information at lower parameter counts. Our approach outperforms traditional feature-extraction registration models that rely on transformers, increasing precision while significantly reducing parameter counts and achieving similar processing speeds to CNN-based models.

E. Network Optimization

In this section, we elaborately explain how to optimize our CHA-Net model. Supervised learning methods require a vast number of labeled data to supervise the training process. However, obtaining ground-truth data for registration is prohibitively expensive, making it challenging to apply such methods in practical settings. As a result, we have chosen unsupervised training as an alternative strategy. Unsupervised models heavily rely on the loss function, which guides the network optimization process. Since there are significant differences in the imaging style among multimodal images, selecting an appropriate loss function becomes crucial, and measuring the similarity between multimodal image pairs is a key challenge. In this article, we adopt the bidirectional-perceptual loss to constrain the alignment of multimodal image pairs

$$L_{\text{sim}} = \| \psi(\phi \circ I_m) - \psi(I_f) \|_1 + \lambda_{\text{rev}} \| \psi(I_m) - \psi(\phi^{-1} \circ I_f) \|_1. \tag{7}$$

The ψ represents the features of the VGG-19 network's Relu3_3 activation, which has the ability to extract content features. The second term leverages registered reversibility not only to maintain alignment between the moving image that has undergone registration and the fixed image but also to restrict the resampling of the fixed image using backward deformation to ensure alignment with the moving image, thus further enhancing the model's robustness. Typically, the distortion in a local region tends to exhibit a similar directionality. To maintain the smoothness of the registered field, a regularization term in the form of a smoothness loss is introduced. It constrains the gradients of the registered field to produce reasonable image distortions

$$L_{\text{smooth}} = \|\text{grad}(\phi)\|_2. \tag{8}$$

Thus, the total loss function for the registration part is computed using the following expression:

$$L_R = L_{\text{sim}} + \lambda_s L_{\text{smooth}} + \lambda_{\text{ha}} L_{\text{ha}}. \tag{9}$$

In addition, in SFD, a discriminator is utilized to enable the model to learn specific style information across multiple modalities. In this article, the Focal Loss is employed as the discriminator loss

$$L_D = -(1 - p_t)^{\gamma} \log(p_t), \quad \gamma = 5. \tag{10}$$

Finally, the total loss function of CHA-Net is represented by

$$L_{\text{total}} = L_R + L_{\text{MI}} + L_D. \tag{11}$$

Algorithm 1 outlines the training procedure in detail.

```
Algorithm 1 CHA-Net Algorithm
Require:
    Fixed images $\{I_f\}$; Moving images $\{I_m\}$;
    Base Encoder $BE$; Content Encoder $CE$; Style Discriminator $D$; transformer Field Estimator $TFE$; Field Calibration Module $FCM_1$, $FCM_2$, $FCM_3$
Ensure:
    A Trained registration network CHA-Net.
while not converged do
    Sample a mini-batch from $\{I_f\}$ and $\{I_m\}$;
    Content Decomposition:
    Compute $\{C_f, S_f\}$ and $\{C_m, S_m\}$ according to Eq. 3;
    Calculate MI loss between $\{C_f, S_f\}$ and $\{C_m, S_m\}$ according to Eq. 4;
    Update $BE$,$CE$,$SD$ according to Eq. 4 and Eq. 10;
    Hierarchical Alignment:
    for $k$ in $1, 2, 3$ do
        Predict $\phi_k$ accroding to Eq. 5;
        Resampler moving feature $M_k$ by $FCM_t$;
    end for
    Update $TFE$, $FCM_1$, $FCM_2$, $FCM_3$, according to Eq. 11
end while
```

III. Experiments

In this section, we provide a brief overview of the dataset, implementation details, and evaluation metrics used in our experiments. Furthermore, to validate the proposed CHA-Net, we conducted extensive parameter experiments, comparative experiments, ablation experiments, efficiency experiments, and visual analyses on multiple datasets.

A. Datasets

1) MS and PAN: The dataset [38] consists of six remote sensing satellite images, including IKONOS, QuickBird, Gaofen-1, WorldView-2, WorldView-3, and WorldView-4. It contains multispectral images and panchromatic images acquired from these satellites. In the multispectral images, the red, green, blue, and near-infrared bands were selected, and together with the panchromatic images, they form the multimodal image pairs. In this dataset, the multispectral images serve as the fixed image, and the panchromatic images serve as the moving image. We randomly selected 618 pairs of 256×256 panchromatic and multispectral images from the dataset, with 495 pairs used for training and the remaining 123 pairs for testing.

2) VIS and NIR: The dataset used in this study comprises multiband images captured by two satellites, namely WorldView-4 and Gaofen-1. These satellites have corresponding spatial resolutions of 1.24 and 8 m, respectively. The multiband images from both satellites consist of four bands, but only the visible and near-infrared bands were selected for the visible and near-infrared dataset, respectively. To simulate real-world scenarios, geometric distortions were introduced into the data in a random manner. In this dataset, the visible light images serve as the fixed image, and the near-infrared

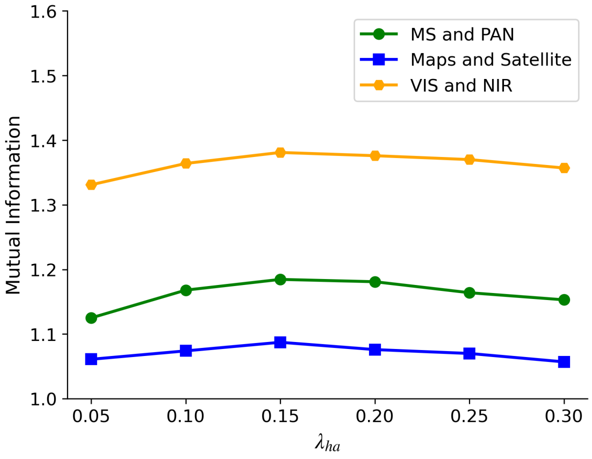


Fig. 2. Evaluation results of λ_{ha} under various weight configurations on three datasets.

TABLE I
KERNEL SIZE ANALYSIS OF CE

Kernel size	MSE↓	LNCC↑	MI↑
$(1, 3, 5)$	0.0012	0.6470	1.0805
$(1, 3, 7)$	0.0015	0.6249	1.0460
$(1, 5, 7)$	0.0017	0.5924	0.9947
$(3, 5, 7)$	0.0014	0.6100	1.0302
$(1, 3, 5, 7)$	0.0012	0.6371	1.0737

images serve as the moving image. The dataset consists of 910 pairs of 256 × 256 image samples, with 728 pairs randomly selected for training and the remaining 182 pairs reserved for testing purposes.

3) Satellite and Maps: This dataset [10] is based on the static Google Maps API and includes multichannel satellite images along with their corresponding static Google Maps images. These images represent the same location, but they are represented in completely different color modes. The dataset comprises image pairs from various periods of the four seasons, making it a diverse collection of images. In this dataset, Google maps serve as the moving image, and satellite images serve as the fixed image. To ensure a fair evaluation of the proposed model, we randomly selected 8822 image pairs from a total of 9710 pairs as the training set, while the remaining 888 pairs were used as the testing set.

B. Implementation Details

To accurately evaluate the model accuracy, the experimental data were pre-registered. To meet the requirements of unaligned image pairs, the data were randomly subjected to geometric distortions before the experiment. We first generated multiple deformation fields by applying different degrees of affine, homography, and elastic transformations to illustrate the registration performance of the method when faced with different types of deformations. Then, we applied these deformation fields to the original images to obtain distorted images [4] The following parameters were used in the experiment: the weight λ_s of the image smoothness loss was set to 10, and the weight coefficient λ_{rev} in the bidirectional-perceptual loss was set to 0.2, by leveraging the experience from previous works [33]. The initial learning rate α was set to 4×10^{-4}. We randomly selected 12 images as a mini-batch and utilized the Adam optimizer (with $\beta_1 = 0.9$ and $\beta_2 = 0.999$) to optimize our model. All experiments were conducted on a computer equipped with an NVIDIA GeForce RTX 2080Ti GPU and an Intel Xeon E5-2630 CPU, using the Pytorch 1.7.1 environment.

C. Evaluation Metrics

To evaluate the model quantitatively, we compared the difference between the resampled image by the model and the ground truth image. We used three common metrics, mean squared error (MSE), local normalized cross correlation (LNCC), and MI, to evaluate the registration results. MSE reflects the overall registration error, while LNCC better reflects the local registration similarity. MI reflects the degree of correlation between the two images. The MI is shown in (4). The equations for MSE and LNCC are as follows:

$$\text{MSE} = \frac{1}{\Omega}\sum_{p=1}^{\Omega}\|X(p) - Y(p)\|_2 \tag{12}$$

$$\text{LNCC} = \sum_{p\in\Omega}\frac{\sum_{p_i}(X(p_i) - \bar{X}(p))((Y(p_i) - \bar{Y}(p))}{\sqrt{\sum_{p_i}(X(p_i) - \bar{X}(p)^2\sum_{p_i}(Y(p_i) - \bar{Y}(p))^2}} \tag{13}$$

where X and Y represent the data distributions of the moved image and the ground truth image, respectively, and p denotes the corresponding pixel of the images.

D. Parameter Experiment

In this section, we evaluated the importance of the hierarchical alignment loss weight coefficient λ_{ha} and the combination of convolutional sizes in the CE of CHA-Net.

The hierarchical alignment loss further constrains the feature distance between multimodal feature spaces, and we changed the value of λ_{ha} to amplify or reduce its constraint on cross-modal feature alignment. A larger value strengthens the constraint on feature distance alignment, while a smaller value relaxes the model's constraint on feature distance alignment.

We set the maximum value of the registration process as the baseline value in the model through multiple experiments, and the experimental results are shown in Fig. 2. The results show that the overall highest accuracy is achieved when λ_{ha} is set to 0.15. This value achieves the optimal constraint on both the overall image alignment and the cross-modal feature distance alignment.

Moreover, to investigate the impact of different combinations of convolutional kernels on the registration performance in the CE, we conducted a series of experiments. We chose Satellite and Maps as the comparison datasets, and the results are presented in Table I. After comprehensive comparisons, we selected the combination of convolutional kernel sizes $(1, 3, 5)$ as the baseline for subsequent experiments, which means that the CE includes 1×1, 3×3, and 5×5 convolutional kernels.

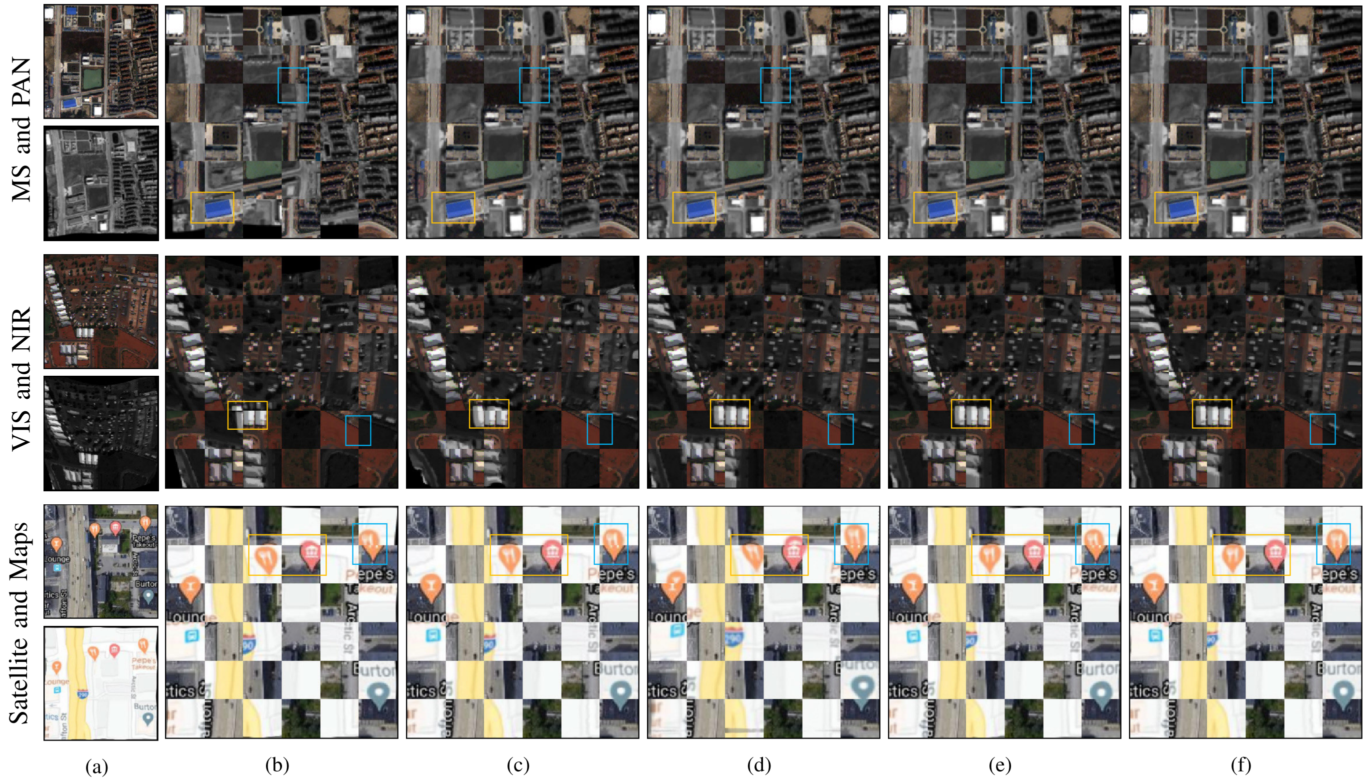


Fig. 3. Comparison of instances from different algorithms on three datasets. (a) Input images, (b) misaligned image, (c) TransMorph result, (d) nemar result, (e) CGRP result, and (f) CHA-Net result. The orange and blue boxes highlighted in the figure, respectively, showcase local regions of interest. By examining the degree of local alignment at the border of the checkboard images within the selected boxes, we compare and assess the performance of these models.

TABLE II
EVALUATION AND COMPARISON OF MENTIONED METHODS

Methods	VIS and NIR			MS and PAN			Satellite and Maps		
	MSE↓	LNCC↑	MI↑	MSE↓	LNCC↑	MI↑	MSE↓	LNCC↑	MI↑
Misaligned Input	0.0045	0.3497	0.8599	0.0265	0.2918	0.6873	0.0154	0.3796	0.7896
SIFT	0.0038	0.4796	0.9845	0.0246	0.4338	0.7528	0.2534	0.1235	0.1621
HOPC	0.0034	0.4966	1.0413	0.0224	0.4590	0.8549	0.0135	0.4444	0.8504
VoxelMorph	0.0024	0.5299	1.1362	0.0151	0.5273	0.9364	0.0094	0.5481	0.9017
CycleMorph	0.0016	0.5541	1.1586	0.0113	0.5695	0.9583	0.0059	0.5678	0.9236
MSDR	0.0017	0.6150	1.2417	0.0114	0.6068	1.0209	0.0366	0.3668	0.7559
NBR-Net	0.0014	0.6112	1.2681	0.0098	0.6002	1.0393	0.0038	0.5815	0.9795
LKU-Net	0.0015	0.6159	1.2757	0.0088	0.6219	1.0377	0.0041	0.5783	0.9947
TransMorph	0.0012	0.6208	1.2850	0.0085	0.6360	1.0456	0.0037	0.5906	1.0073
NeMAR	0.0022	0.6259	1.2928	0.0091	0.6632	1.0346	0.0194	0.1889	0.3880
CGRP	0.0008	0.6442	1.3404	0.0089	0.6734	1.0926	0.0980	0.5402	0.9182
Ours	0.0007	0.6748	1.4098	0.0071	0.7196	1.1846	0.0012	0.6470	1.0805

E. Accuracy Analysis

In this section, we compare the performance of CHA-Net with other approaches to evaluate its efficacy on multimodal datasets. The multimodal datasets introduced are chosen as the evaluation dataset, consisting of image pairs with varying degrees of geometric distortion in the moving image to the fixed image. We conduct both quantitative and qualitative analyses on the methods SIFT, HOPC [22], VoxelMorph [39], CycleMorph [40], MSDR, NBR-Net [4], [15], LKU-Net [24], TransMorph [27], NeMAR [34], CGRP [33], and CHA-Net across three distinct datasets to assess their performance.

1) Quantitative Accuracy Analysis: In this section, we present a quantitative evaluation of different methods on three multimodal datasets using the metrics MSE, LNCC, and MI. Table II illustrates the experimental results of each method on these three metrics. All experiments were conducted using the multimodal version of the implementation. Among these results, CHA-Net achieved the highest accuracy across all datasets and exhibited consistent registration performance despite changes in modality. SIFT is capable of capturing some nonrigid distortions when dealing with datasets with small style variations. However, when faced with the satellite and maps dataset, it produces completely distorted images.

TABLE III
RESULTS OF ABLATION EXPERIMENTS

Methods	BE	HRA	SFD	MSE↓	LNCC↑	MI↑
CHA-Net w/o HRA & SFD	✓			0.0067	0.5402	0.9373
CHA-Net w/o SFD	✓	✓		0.0049	0.5852	0.9861
CHA-Net w/o HRA	✓		✓	0.0032	0.5988	1.0130
CHA-Net	✓	✓	✓	0.0012	0.6470	1.0805

In contrast, HOPC, utilizing structural similarity descriptors, demonstrates robust performance across all datasets. NBR-Net, with its registration cycle consistency, maintains good registration performance and robustness when confronted with three different datasets. TransMorph and LKU-Net also demonstrated stable performance on each dataset, with relatively good registration results. MSDR employed a multistep iterative approach to perform multiple corrections on the images, achieving good performance on the MS and PAN datasets as well as the VIS and NIR datasets. However, when faced with the Satellite and Maps dataset, the model failed to converge to a correct result due to the significant differences in image styles. Consequently, erroneous registration results were produced. CGRP and NeMAR employ image translation to generate cross-modal images, which heavily rely on the performance of the image translation network. Although they demonstrated relatively good results on the MS and PAN datasets as well as the VIS and NIR datasets, when faced with the significantly larger modal gap present in the Satellite and Maps dataset, these methods failed to train a suitable image translation network, consequently producing erroneous registration results. Although CycleMorph and VoxelMorph may not exhibit exceptional accuracy in registration, they can still produce relatively stable results across various datasets.

Experimental results demonstrate that CHA-Net is a versatile model that can be applied to various types of multimodal image registration tasks. CHA-Net outperforms existing registration models with its comprehensive and accurate registration results.

2) Qualitative Accuracy Analysis: Fig. 3 illustrates multiple registration examples across three multimodal datasets. We combine the two images in a checkerboard pattern for better visualization. Due to space limitations, we have selected algorithms that are currently performing well and representative for visual comparison to showcase the registration accuracy. Generally, CHA-Net showcases robust and desirable geometric correction of moving images across all datasets, compared to the fixed images. It presents excellent alignment performance while preserving clear structures, showcasing optimal registration accuracy. Furthermore, TransMorph displays certain degrees of geometric correction to the image distortions across these datasets, yet some misalignments remain unresolved. The Nemar and CGRP methods can achieve reasonable correction on the datasets MS and PAN, as well as VIS and NIR. However, as the difficulty of registering multimodal datasets increases, these methods fail to generate reasonable pseudo-images for image registration based on multimodal conversion in the satellite and maps dataset, resulting in erroneous image alignment.

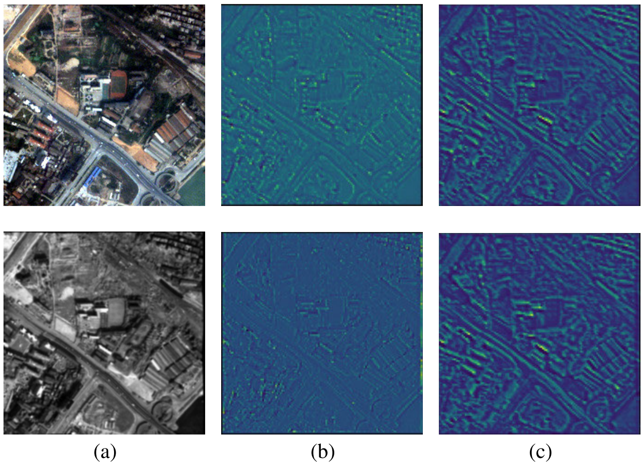


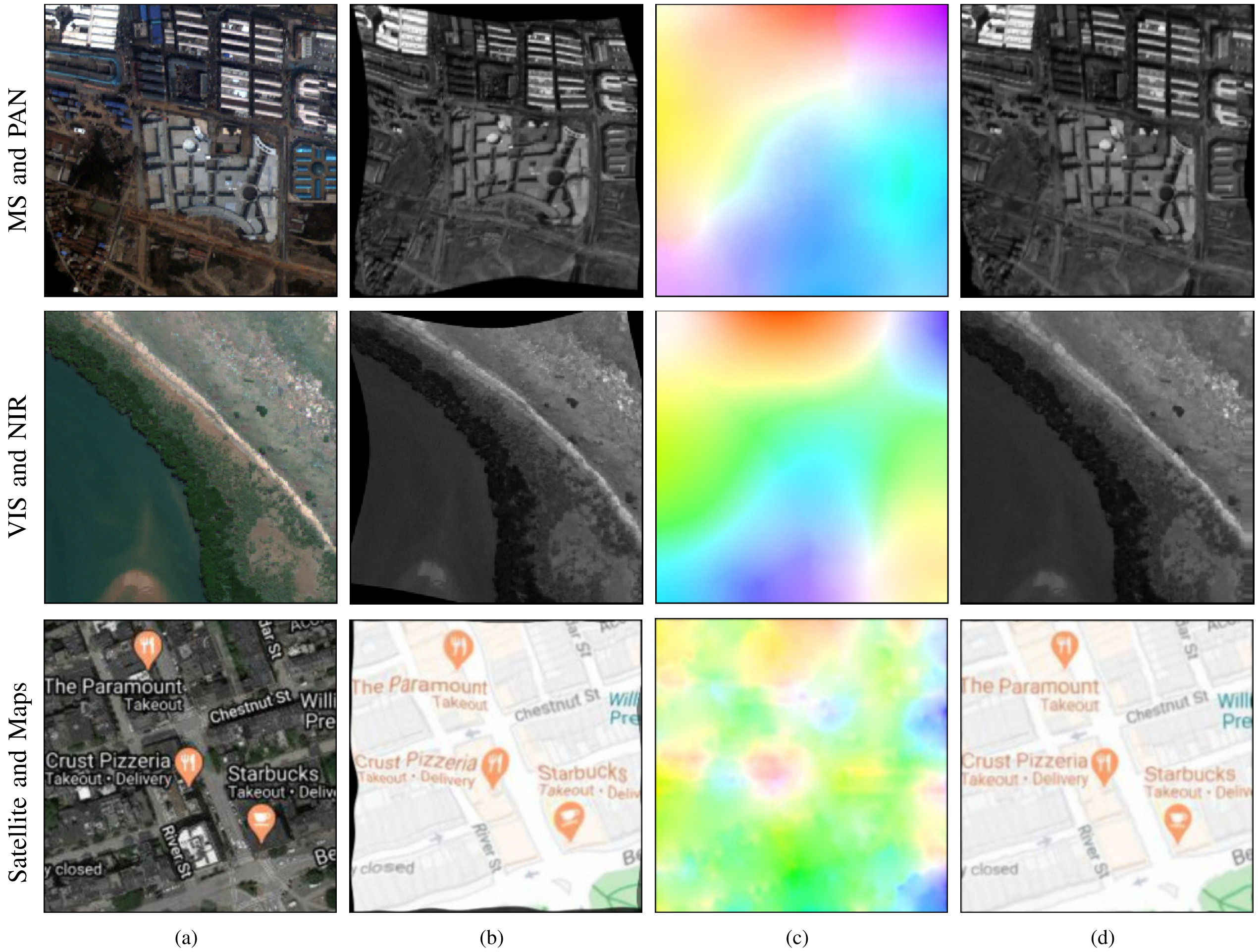
Fig. 4. Visualization of the decoupled feature maps extracted by SFD, with the first row representing the fixed image and the second row representing the moving image. Specifically, (a) input data, (b) feature maps extracted by the Siamese neural network, and (c) feature maps acquired via SFD. It is apparent that compared to column (b), column (c) exhibits a greater focus on the shared information between the two modalities, showcasing the effectiveness of SFD in capturing Sharing information.

F. Ablation Study

Continuing our investigation through experimentation in this section, we conducted an ablation study on the satellite and maps dataset to evaluate the effectiveness of the critical components of the proposed network. Specifically, we constructed various model configurations by substituting SFD and HRA with different alternatives, to observe the effect of SFD and HRA in the models. The experimental results are presented in Table III.

1) Effectiveness of SFD: In the preceding section, we proposed the SFD structure, which employs a discriminator to decompose the image into two mutually orthogonal feature vectors. This, in turn, controls the model's focus on content information for registration purposes. In order to ascertain the impact of SFD on the registration process, we replaced SFD in the original model with only the Siamese neural network, serving as the feature extraction component. The accuracy was then evaluated across three datasets, the results of which are detailed in Table III.

As indicated in Table III, the incorporation of SFD generates noteworthy improvements in the satellite and maps dataset



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

Fig. 5. Visualization of registration results (a) fixed image, (b) moving image, (c) registration field, and (d) moved image. Column (c) shows the visualization of the registration field using the standard optical-flow visualization. Areas with distinct color differences from their surroundings highlight the presence of significant geometric distortions.

datasets. SFD serves as a proficient solution for handling multimodal issues by regulating the model's focus on content information and minimizing the feature distribution disparities between the various modalities.

To further validate the impact of domain adaptation on the model's performance, we visualized the extracted feature maps during the registration process. Specifically, we selected a pair of images consisting of a multispectral image and a panchromatic image. We averaged the feature maps in the network to obtain single-channel images for visualization. This was done to compare the visual effects before and after using SFD. As showcased in Fig. 4, after performing distribution alignment with SFD, the features encoded by our model exhibit increased structural and textural coherence, with fewer underlying stylistic variations. This renders our model more adept at tackling multimodal image pairs and enhances its ability to achieve accurate registration results.

2) Effectiveness of HRA: To better capture nonrigid distortions, we proposed the HRA structure in Section II-D. By enhancing the model's focus on different scales and gradually correcting the registration field with multilevel features, this structure enables pixel-level registration.

To investigate the effectiveness of HRA, we replaced its role in the model with the one-step generation of the registration field using CNNs.

From Table III, it is evident that the incorporation of HRA leads to a noticeable enhancement of the model's performance on the satellite and maps dataset datasets. By leveraging HRA, the model directs its attention to multilevel features, enabling it to capture distortions of varying scales. This proves to be an effective solution for addressing the geometric distortions inherent in nonrigid registration.

G. Futher Analysis

1) Visualization Experiment: To further evaluate the registration performance of CHA-Net, we employed three datasets as mentioned in the previous section for visualization purposes. Moreover, the generated registration fields during the registration process were further visualized using the standard optical-flow visualization [41], as depicted in Fig. 5. Areas with distinct color differences from their surroundings highlight the presence of significant geometric distortions. It is evident that the moving image exhibits a substantial amount of geometric distortion compared to the fixed image. However, after being registered by CHA-Net, the moved image aligns correctly with the fixed image in terms of its content, and the registration field effectively captures the distortions present in the moving image.

2) Fusion Experiments: To further validate the effectiveness of the registration methods, we evaluated their performance in

TABLE IV
FUSION RESULTS OF VARIOUS REGISTRATION METHODS

Methods	SAM↓	ERGAS↓	SSIM↑
Misaligned Input	2.1957	7.8591	0.4341
SIFT	2.1884	7.5094	0.4651
HOPC	1.9982	6.6342	0.4897
VoxelMorph	2.0063	6.6459	0.4937
CycleMorph	1.9754	6.5686	0.5141
MSDR	1.9318	6.3475	0.5701
NBR-Net	1.8802	5.8723	0.5822
LKU-Net	1.8941	6.0121	0.5762
TransMorph	1.8998	5.8457	0.6194
NeMAR	1.9247	5.7871	0.5912
CGRP	1.8701	5.7002	0.6245
CHA-Net	1.8372	5.5740	0.6333

image registration by comparing the fusion effects achieved by different registration methods. In our experiments, we utilized UCGAN [42] as the image fusion method and employed evaluation metrics based on the Wald protocol [43], which included spectral angle mapper (SAM), relative dimensionless global error in synthesis (ERGAS), and SSIM. We selected the MS and PAN datasets for our experiments. In the image fusion experiments, we initially conducted image registration using various registration methods on the remote sensing images. Following that, we employed UCGAN to execute the fusion process on the registered images. Furthermore, we conducted a quantitative analysis of the fusion results using the evaluation metrics mentioned earlier. The experimental results are presented in Table IV, where it can be observed that CHA-Net achieved the best performance, demonstrating its effectiveness in image fusion.

3) Efficiency Experiments: Despite its superior performance in various visual tasks, transformer models are known for their computational resource requirements, which cannot be overlooked in comparison to traditional CNN models. In HRA, we leverage the transformer model for global alignment estimation. In this section, we conduct an efficiency analysis and comparison of CHA-Net with other state-of-the-art registration models, in terms of their parameters and speed. Specifically, we conducted experiments on a pair of images with 256 × 256 resolution as input.

The results are presented in Table V. TransMorph, which relies on transformers as the encoder, exhibits the highest resource consumption and significantly slower speed. In comparison, the CNN-based model LKU-Net achieves a significant improvement in both speed and resource consumption while sacrificing a relatively small amount of accuracy. Nemar and CGRP both utilize generative models as crucial components for registration, resulting in larger parameter sizes and higher resource consumption. However, Nemar only employs a registration network with fewer parameters during the inference process, leading to better performance in terms of speed. It should be noted that CGRP is not an end-to-end network, so its speed is not analyzed in this comparison.

TABLE V
EFFICIENCY ANALYSIS OF CHA-NET

Methods	LKU-Net	TransMorph	Nemar	CGRP	CHA-Net
Parm. (M)	1.553	31.014	16.205	28.272	7.136
FPS	38.26	12.50	34.00	/	29.64

Overall, CHA-Net has significantly fewer parameters than models based on transformer or GAN, with a parameter count below 10 M. In terms of speed, CHA-Net leverages transformers to model low-resolution features, enabling it to achieve a speed of up to 29.64 FPS, which approaches real-time performance for image registration.

IV. CONCLUSION

In this article, we proposed CHA-Net, a nonrigid multimodal registration network designed for remote sensing images. CHA-Net corrects the unaligned areas between multimodal image pairs with nonrigid distortion in an unsupervised manner. By utilizing domain-adaptive feature decoupling, CHA-Net reduces the feature distance between the two modalities. It employs the Siamese neural network to extract common information from the multimodal image pairs and filters out noise, allowing the model to focus on content features and possess a robust resistance to nonlinear radiation of multimodal images. Furthermore, our proposed HRA structure retains the local modeling capability while enhancing the model's ability to capture geometric distortions on a global scale. We also designed the hierarchical alignment loss to achieve spatial alignment at different scales of features. CHA-Net is suitable for various types of multimodal image registration tasks. We conducted experiments on datasets of various types and compared CHA-Net to state-of-the-art models. The results showcase that CHA-Net exhibits higher performance and strong robustness. However, in our research, when dealing with the nonrigid geometric distortions between optical and SAR images, we have observed that while CHA-Net is capable of capturing some fundamental geometric distortions, it still encounters limitations in certain cases. This can be attributed to the presence of nonlinear radiation and a significant amount of noise between SAR and optical images, which adversely affects the nonrigid registration process. Consequently, our future research will be dedicated to a more in-depth investigation of nonrigid registration techniques specifically tailored for optical and SAR image pairs.

REFERENCES

[1] X. Jiang, J. Ma, G. Xiao, Z. Shao, and X. Guo, "A review of multimodal image matching: Methods and applications," *Inf. Fusion*, vol. 73, pp. 22–71, Sep. 2021.

[2] X. Zhang, W. Huang, Q. Wang, and X. Li, "SSR-NET: Spatial–spectral reconstruction network for hyperspectral and multispectral image fusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 7, pp. 5953–5965, Jul. 2021.

[3] R. Dian, S. Li, L. Fang, and Q. Wei, "Multispectral and hyperspectral image fusion with spatial–spectral sparse representation," *Inf. Fusion*, vol. 49, pp. 262–270, Sep. 2019.

[4] Y. Xu, J. Li, C. Du, and H. Chen, "NBR-net: A nonrigid bidirectional registration network for multitemporal remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5620715.

[5] H. Zhang and P. Ren, "Game theoretic hypergraph matching for multi-source image correspondences," *Pattern Recognit. Lett.*, vol. 87, pp. 87–95, Feb. 2017.

[6] A. Zampieri, G. Charpiat, N. Girard, and Y. Tarabalka, "Multimodal image alignment through a multiscale chain of neural networks with application to remote sensing," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Munich, Germany, Sep. 2018, pp. 1–17.

[7] N. Girard, G. Charpiat, and Y. Tarabalka, "Noisy supervision for correcting misaligned cadaster maps without perfect ground truth data," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Valencia, Spain, Jul. 2019, pp. 10103–10106.

[8] Y. Xiang, R. Tao, F. Wang, H. You, and B. Han, "Automatic registration of optical and SAR images via improved phase congruency model," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 5847–5861, 2020.

[9] W. Lee, D. Sim, and S.-J. Oh, "A CNN-based high-accuracy registration for remote sensing images," *Remote Sens.*, vol. 13, no. 8, p. 1482, Apr. 2021.

[10] Y. Zhao, X. Huang, and Z. Zhang, "Deep Lucas-Kanade homography for multimodal image alignment," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 15945–15954.

[11] M. C. Hansen et al., "High-resolution global maps of 21st-century forest cover change," *Science*, vol. 342, no. 6160, pp. 850–853, Nov. 2013.

[12] G. Máttyus, S. Wang, S. Fidler, and R. Urtasun, "HD maps: Fine-grained road segmentation by parsing ground and aerial images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3611–3619.

[13] Y. Ye, B. Zhu, T. Tang, C. Yang, Q. Xu, and G. Zhang, "A robust multimodal remote sensing image registration method and system using steerable filters with first- and second-order gradients," *ISPRS J. Photogramm. Remote Sens.*, vol. 188, pp. 331–350, Jun. 2022.

[14] Y. Ye, T. Tang, B. Zhu, C. Yang, B. Li, and S. Hao, "A multiscale framework with unsupervised learning for remote sensing image registration," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5622215.

[15] M. Papadomanolaki, S. Christodoulidis, K. Karantzalos, and M. Vakalopoulou, "Unsupervised multistep deformable registration of remote sensing imagery based on deep learning," *Remote Sens.*, vol. 13, no. 7, p. 1294, Mar. 2021.

[16] Y. Xiang, N. Jiao, F. Wang, and H. You, "A robust two-stage registration algorithm for large optical and SAR images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5218615.

[17] B. Zhu, Y. Ye, L. Zhou, Z. Li, and G. Yin, "Robust registration of aerial images and LiDAR data using spatial constraints and Gabor structural features," *ISPRS J. Photogramm. Remote Sens.*, vol. 181, pp. 129–147, Nov. 2021.

[18] D. Quan et al., "Deep feature correlation learning for multi-modal remote sensing image registration," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4708216.

[19] B. Zhu, C. Yang, J. Dai, J. Fan, Y. Qin, and Y. Ye, "R2FD2: Fast and robust matching of multimodal remote sensing images via repeatable feature detector and rotation-invariant feature descriptor," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5606115.

[20] G. Lv, Q. Chi, M. Awrangjeb, and J. Li, "Robust registration of multispectral satellite images based on structural and geometrical similarity," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2021, Art. no. 5001705.

[21] L. Zhou, Y. Ye, T. Tang, K. Nan, and Y. Qin, "Robust matching for SAR and optical images using multiscale convolutional gradient features," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 4017605.

[22] Y. Ye, J. Shan, L. Bruzzone, and L. Shen, "Robust registration of multimodal remote sensing images based on structural similarity," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 5, pp. 2941–2958, May 2017.

[23] S. Wang, D. Quan, X. Liang, M. Ning, Y. Guo, and L. Jiao, "A deep learning framework for remote sensing image registration," *ISPRS J. Photogramm. Remote Sens.*, vol. 145, pp. 148–164, Nov. 2018.

[24] X. Jia, J. Bartlett, T. Zhang, W. Lu, Z. Qiu, and J. Duan, "U-net vs transformer: Is U-net outdated in medical image registration?" in *Proc. Mach. Learn. Med. Imag. Int. Workshop Mach. Learn. Med. Imag.*, Singapore, Cham, Switzerland: Springer, Sep. 2022, pp. 151–160.

[25] L. Zeng, Y. Du, H. Lin, J. Wang, J. Yin, and J. Yang, "A novel region-based image registration method for multisource remote sensing images via CNN," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 1821–1831, 2021.

[26] S. Hoffmann, C.-A. Brust, M. Shadaydeh, and J. Denzler, "Registration of high resolution SAR and optical satellite imagery using fully convolutional networks," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2019, pp. 5152–5155.

[27] J. Chen, E. C. Frey, Y. He, W. P. Segars, Y. Li, and Y. Du, "TransMorph: Transformer for unsupervised medical image registration," *Med. Image Anal.*, vol. 82, Nov. 2022, Art. no. 102615.

[28] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9992–10002.

[29] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2242–2251.

[30] J. Zhang, W. Ma, Y. Wu, and L. Jiao, "Multimodal remote sensing image registration based on image transfer and local features," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 8, pp. 1210–1214, Aug. 2019.

[31] S.-Y. Cao, H.-L. Shen, S.-J. Chen, and C. Li, "Boosting structure consistency for multispectral and multimodal image registration," *IEEE Trans. Image Process.*, vol. 29, pp. 5147–5162, 2020.

[32] W. Ma, J. Zhang, Y. Wu, L. Jiao, H. Zhu, and W. Zhao, "A novel two-step registration method for remote sensing images based on deep and local features," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 7, pp. 4834–4843, Jul. 2019.

[33] W. Di, L. Jinyuan, F. Xin, and R. Liu, "Unsupervised misaligned infrared and visible image fusion via cross-modality image generation and registration," in *Proc. Int. Joint Conf. Artif. Intell. (IJCAI)*, 2022, pp. 1–9.

[34] M. Arar, Y. Ginger, D. Danon, A. H. Bermano, and D. Cohen-Or, "Unsupervised multi-modal image registration via geometry preserving image-to-image translation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 13407–13416.

[35] L. Wu, M. Lu, and L. Fang, "Deep covariance alignment for domain adaptive remote sensing image segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5620811.

[36] A. Wu, R. Liu, Y. Han, L. Zhu, and Y. Yang, "Vector-decomposed disentanglement for domain-invariant object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9322–9331.

[37] M. Jaderberg et al., "Spatial transformer networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 28, 2015, pp. 1–11.

[38] X. Meng et al., "A large-scale benchmark data set for evaluating pansharpening performance: Overview and implementation," *IEEE Geosci. Remote Sens. Mag. Replaces Newsletter*, vol. 9, no. 1, pp. 18–52, Mar. 2021.

[39] G. Balakrishnan, A. Zhao, M. R. Sabuncu, A. V. Dalca, and J. Guttag, "An unsupervised learning model for deformable medical image registration," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 9252–9260.

[40] B. Kim, D. H. Kim, S. H. Park, J. Kim, J.-G. Lee, and J. C. Ye, "CycleMorph: Cycle consistent unsupervised deformable image registration," *Med. Image Anal.*, vol. 71, Jul. 2021, Art. no. 102036.

[41] S. Baker, D. Scharstein, J. P. Lewis, S. Roth, M. J. Black, and R. Szeliski, "A database and evaluation methodology for optical flow," *Int. J. Comput. Vis.*, vol. 92, no. 1, pp. 1–31, Mar. 2011.

[42] H. Zhou, Q. Liu, D. Weng, and Y. Wang, "Unsupervised cycle-consistent generative adversarial networks for pan sharpening," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5408814.

[43] L. Wald, T. Ranchin, and M. Mangolini, "Fusion of satellite images of different spatial resolutions: Assessing the quality of resulting images," *Photogramm. Eng. Remote Sens.*, vol. 63, no. 6, pp. 691–699, 1997.