A Multiscale Framework With Unsupervised Learning for Remote Sensing Image Registration

Yuanxin Ye, *Member, IEEE*, Tengfeng Tang, Bai Zhu, Chao Yang, Bo Li, and Siyuan Hao, *Member, IEEE*

***Abstract*— Registration for multisensor or multimodal image pairs with a large degree of distortions is a fundamental task for many remote sensing applications. To achieve accurate and low-cost remote sensing image registration, we propose a multiscale framework with unsupervised learning, named MU-Net. Without costly ground truth labels, MU-Net directly learns the end-to-end mapping from the image pairs to their transformation parameters. MU-Net stacks several deep neural network (DNN) models on multiple scales to generate a coarse-to-fine registration pipeline, which prevents the backpropagation from falling into a local extremum and resists significant image distortions. We design a novel loss function paradigm based on structural similarity, which makes MU-Net suitable for various types of multimodal images. MU-Net is compared with traditional feature-based and area-based methods, as well as supervised and other unsupervised learning methods on the optical-optical, optical-infrared, optical-synthetic aperture radar (SAR), and optical-map datasets. Experimental results show that MU-Net achieves more comprehensive and accurate registration performance between these image pairs with geometric and radiometric distortions. We share the code implemented by Pytorch at https://github.com/yeyuanxin110/MU-Net.**

***Index Terms*— Image registration, multimodal images, multiscale framework, unsupervised learning.**

I. Introduction

REMOTE sensing image registration (RSIR) aims to identify and correspond the same or similar structure or content from two or more images at the pixel level, and these images are captured at different times, by diverse satellite sensors, or even from different viewpoints [1]. RSIR directly influences the performance of the following tasks, such as image fusion, change detection, deformation monitoring, and land-use analysis [2]. Thus, RSIR especially multimodal RSIR is a necessary and preliminary task.

Manuscript received November 23, 2021; revised February 7, 2022 and March 23, 2022; accepted April 10, 2022. Date of publication April 18, 2022; date of current version May 5, 2022. This work was supported by the National Natural Science Foundation of China under Grant 41401369 and Grant 62171247. *(Corresponding author: Siyuan Hao.)*

Yuanxin Ye, Tengfeng Tang, Bai Zhu, and Chao Yang are with the Faculty of Geosciences and Environmental Engineering, Southwest Jiaotong University, Chengdu 610031, China, and also with the State-Province Joint Engineering Laboratory of Spatial Information Technology for High-Speed Railway Safety, Southwest Jiaotong University, Chengdu 611756, China (e-mail: yeyuanxin@home.swjtu.edu.cn).

Bo Li is with the School of Electronics and Information, Northwestern Polytechnical University, Xi'an 710129, China (e-mail: libo@nwpu.edu.cn).

Siyuan Hao is with the School of Information and Control Engineering, Qingdao University of Technology, Qingdao 266033, China (e-mail: lemonbananan@163.com).

Digital Object Identifier 10.1109/TGRS.2022.3167644

Multimodal RSIR mainly includes image registration among optical, infrared, synthetic aperture radar (SAR), multispectral, and other data. The optical images have rich texture details and high resolution. However, they are greatly affected by weather conditions; for example, it is difficult to obtain the information of the ground surface at night or when obscured by the clouds. The infrared images are acquired according to the thermal radiation characteristics, which reflect the existence and position of the target, but the edge of the target is relatively blurred and the resolution is low. The SAR is an active earth observation microwave system, which can image all day and all weather, and penetrate clouds and fog. Therefore, these multimodal images have good complementarity. The features of multimodal images are organically integrated through image fusion, which can fully utilize the potential of various data and provide the accuracy and efficiency of image interpretation and information extraction. The main challenge of multimodal RSIR is significant radiometric differences and geometric distortions, which makes conventional similarity measures or feature descriptors based on image intensity information almost unusable.

From traditional to deep learning (DL) techniques, many inspiring ideas and methods for image registration have been developed in the fields of computer vision, medical imaging, or remote sensing. In past decade, traditional methods can be generally classified into two categories: feature-based methods [3] and area-based methods [4]. Through these methods, the correspondence between images is found; then, the transformation parameters (TPs), including affine TPs [5], homography TPs [6], and deformation vector field (DVF) [7], are fitted by a robust estimation algorithm like random sample consensus (RANSAC) [8] or marginalizing sample consensus (MAGSAC) [9].

Feature-based methods extract salient features (e.g., points, lines, and regions) from two images and establish their correspondences. Many representative feature-based methods, such as the scale-invariant feature transform (SIFT) [10], the speed-up robust feature (SURF) [11], and the oriented FAST and rotated BRIEF (ORB) [12], are suitable for single-modal images with rotation and scale differences due to viewpoint or time changes, but vulnerable to multimodal image matching. This is because they are mainly dependent on detecting highly repeatable features, which are often influenced by large intensity or texture differences between multimodal images. As a result, their matching performance is sensitive to feature differences. It is commonly not repeated for the features that

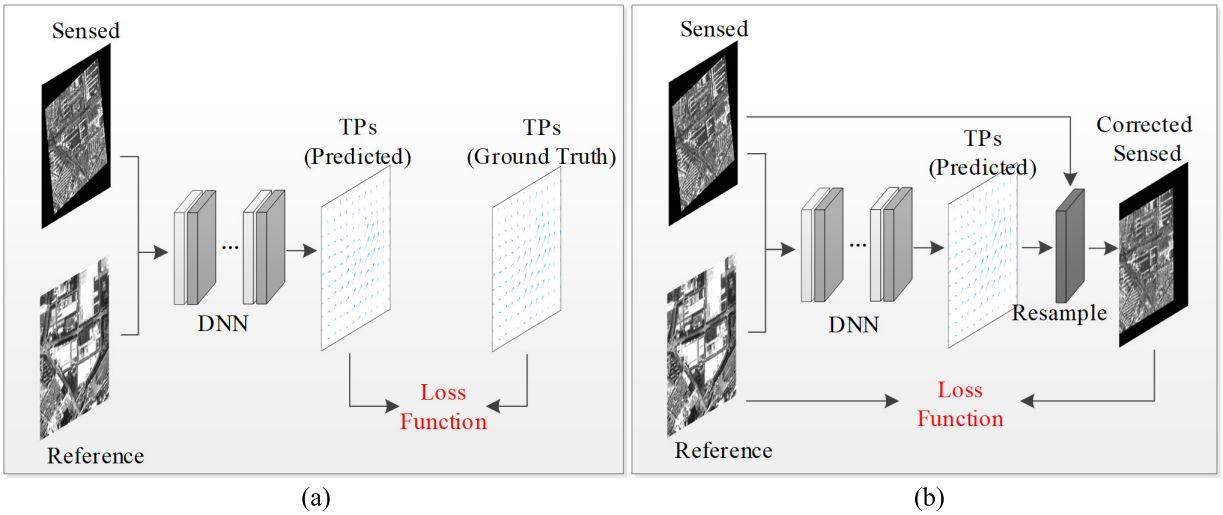

Fig. 1. General architecture of: (a) supervised methods and (b) unsupervised methods for end-to-end image registration.

extracted from multimodal images due to large differences on intensity or texture [13], which greatly influences the matching accuracy. Thus, in RSIR, feature-based methods are difficult to extract common features between multimodal images such as optical and SAR images, which have significant nonlinear radiometric differences [14].

For area-based methods, the similarity of image patch pairs is commonly evaluated by a template matching scheme, and the central points of the patch pairs with the best similarity are identified as the correspondences. When structure features are used for similarity evaluation, area-based methods can effectively address nonlinear radiometric differences between multimodal images [15]. Furthermore, the matching operation can be performed in the frequency to speed up the calculation [16]. However, area-based methods usually cannot effectively handle the images with large geometrical distortions [17].

The two types of traditional methods, respectively, have the abovementioned problems. In addition, both methods commonly include a matching process which integrating features or local descriptors extracted by handcraft rather than automated learning. Because of no information feedback among feature extraction, description, and matching, these approaches lack deep-level semantic information and have limited applicability that can only be applied on some specific cases [18]. When image data source changes, these handcrafted features usually need redesigning to maintain the matching performance. Therefore, traditional methods are often difficult to handle both geometric distortions and radiometric differences between multimodal images.

With the rapid development of artificial intelligence techniques in recent years, a growing number of researches focus on DL. To a certain extent, DL methods can solve the abovementioned shortcomings of traditional methods and have been increasingly applied in remote sensing field, such as image retrieval [19], [20], image classification [21], [22], and object detection [23], [24]. This is because of their completely data-driven schemes, which learn the deep-level information from data automatically without manually setting in feature learning, trying to abstract the distribution structure from the input data. In general, DL methods can be classified into two categories: 1) integrated [25] and 2) end-to-end [26]–[28].

Integrated learning methods usually integrate a deep neural network (DNN) [29] into the traditional feature-based or area-based methods and extract distinctive feature descriptors from the auto-learned feature maps with deep-level semantic information. The idea is that the original images are input into a designed DNN to generate deep feature maps, and traditional operations like keypoint detection, feature description, template sliding, or feature matching are conducted on the auto-learned deep feature maps instead of the original images or the handicraft feature maps. These integrated methods based on DNN can make feature detection or description auto-learned. However, they still require designing a specific DNN for different data, which is not universal. Compared with traditional methods, the calculated amount increases many times, but the registration effect has not been improved significantly.

End-to-end learning methods aim to directly predict the TPs. The frameworks of end-to-end methods commonly consist of a DNN with an unrestricted form to extract deep-level semantic information and predict the TPs; a spatial transformer network (STN) [30] is used to perform image registration, and an optimizer is used to backpropagate the DNN. According to whether the optimizer needs the ground truth TPs, end-to-end learning methods can be divided into supervised end-to-end learning methods (hereafter called supervised learning methods) and unsupervised end-to-end learning methods (hereafter called unsupervised learning methods) [31], and their common architectures are shown in Fig. 1, respectively.

For supervised learning methods, as shown in Fig. 1(a), the loss function is commonly designed based on minimizing the discrepancy between the predicted TPs and the ground truth TPs in the process of backpropagation. Considering the characteristics of supervised methods themselves, whose model requires fitting the entire network containing thousands or even

millions of parameters. To prevent overfitting, the model needs training by a large number of image pairs with ground truth TPs. However, one big challenge is that a number of ground truth TPs are costly and hard to acquire for real images in RSIR [32]. Such limitation makes supervised learning methods difficult to be widely applied in practice.

As Fig. 1(b) shown for unsupervised learning methods, ground truth TPs are not required. In the process of backpropagation, unsupervised methods generally optimize the similarity between images [33]. In recent years, such methods have been gradually developed and applied in medical image registration because they solve the problem that the model cannot be trained effectively with few numbers of trainable data and no ground truth TPs [27], [28]. However, it may not be appropriate to directly apply related methods to RSIR for the following reasons. First, current methods cannot effectively handle noise and nonlinear radiometric differences, which make these methods vulnerable for multimodal RSIR. Second, these methods usually require medical images to be affinely prealigned before training, whereas in RSIR, it is the goal that making images affine alignment. Current methods can effectively address geometric deformation to some degree, but when image has significant geometric and radiometric differences, these methods often suffer large registration errors. Nevertheless, the unsupervised learning methods of medical images still provide enlightening ideas for our proposed framework.

In general, current DL methods cannot effectively handle large geometric distortions and radiometric differences between remote sensing images without the ground truth TPs. To address the issue, we propose a multiscale framework with unsupervised learning for RSIR, named MU-Net, and it is an end-to-end mapping scheme from the input image pairs to their TPs. We stack several DNN models for a coarse-to-fine registration pipeline, and each DNN model represents a workflow performed on an individual scale. On each scale, the corresponding DNN is trained by optimizing the similarity of image pairs, thus circumventing the need for a large number of trainable data and ground truth TPs. First, each DNN model on its scale is individually and successively trained to initialize the network weights. Second, the above models are stacked in a cascading way to form a combined registration pipeline, and the parameters of which are jointly trained to output TPs with fine accuracy. In addition, the similarity evaluation of image pairs is performed on structural features rather than image intensity, which is suitable for multimodal RSIR. Moreover, we also find that adding deep residual (DR) convolution and channel attention mechanisms in DNN can enhance the framework's robustness to image geometric distortions and radiometric differences.

In general, our main contributions have three aspects.

1) We propose a registration framework (named MU-Net) with unsupervised learning, which is an end-to-end mapping scheme from the image pairs to their TPs.
2) We stack several DNN models on multiple scales to generate a coarse-to-fine registration pipeline, which avoids being trapped in a local extremum and resists a large range of image distortions. Moreover, each DNN employs DR convolution and channel attention mechanisms to improve its robustness.
3) We design a novel loss function paradigm based on structural similarity, which makes MU-Net suitable for various types of multimodal images.

II. Related Works

A. Traditional Image Registration

In this section, we briefly review traditional image registration methods into two categories: feature-based methods and area-based methods.

1) Feature-Based Methods: Feature-based methods mainly extract salient features and use the similarity of these features for image matching. As a representative, SIFT [10] has pushed feature-based methods to a new level, and many algorithms are proposed on the basis of SIFT's idea, such as SURF [11] and the uniform robust scale-invariant feature transform (UR-SIFT) [34]. These SIFT-like methods are suitable for the matching of single-modal images, whereas not appropriate for multimodal images with nonlinear radiometric differences such as optical-SAR image pairs. To address that, Ye and Shen [35] proposed a local feature descriptor named local histogram of oriented phase congruency (LHOPC), in which phase congruency is used instead of intensity or gradients to generate an oriented histogram representation. Xiang *et al.* [36] adopted diverse operators to evaluate gradients of optical-SAR image pairs, constructed scale space, and proposed a novel strategy of keypoints matching. Li *et al.* [17] proposed a feature registration algorithm named radiation-variation insensitive feature transform (RIFT), which introduced a maximum index map instead of gradients to describe the local feature. Overall, the main challenge of feature methods is to extract highly repeatable and distinct features and match them correctly [1].

2) Area-Based Methods: Area-based methods often detect correspondences by evaluating the similarity of images. Some widely used similarity metrics are the sum of squared difference (SSD), the normalized cross correlation (NCC) [37], and the mutual information (MI) [38]. Performance of SSD or NCC is easily affected by significant nonlinear radiometric changes. Although MI can address radiometric changes to a certain extent, it does not consider the spatial relationship of neighborhood pixels, which causes a decline in the matching performance. For multimodal RSIR tasks, there are several recent studies that improve the effectiveness of similarity metrics. First, high-level feature descriptors are integrated into similarity metrics to cope with different modalities. By combing the NCC metric and the histogram of oriented phase congruency (HOPC) feature descriptor, a similarity metric named HOPC_{ncc} [39] is designed to detect correspondences by using a template matching scheme, which makes the traditional similarity metric (i.e., NCC) suitable for optical-SAR image registration. Second, similarity metrics based on information theory have been extensively applied in multimodal image registration. In order to address the limitation of MI when dealing with insufficient scene overlap image pairs, Xu *et al.* [40] adopted Jeffrey's divergence as a novel similarity metric for template matching. Third, a fast way to perform

template matching is to evaluate the similarity in frequency domain. By adopting the fast Fourier transform (FFT) [16], Ye *et al.* [14] transformed local feature representation into frequency domain, which vastly accelerates the calculation of similarity evaluation. In general, a robust similarity metric and its computational efficiency play pivotal roles on area-based methods.

B. DL Methods for Image Registration

In this section, we briefly review the DL methods for image registration developed in recent years, including integrated methods and end-to-end methods of supervision and unsupervision.

1) Integrated Methods: Integrated learning methods usually integrate a DNN into traditional feature-based or area-based methods and generate discriminative feature representations from auto-learned feature maps with deep-level semantic information. Yang *et al.* [25] adopted a DNN to obtain local feature descriptors for SIFT keypoints instead of using the handcrafted manner and introduced a strategy of dynamic point selection to promote the matching performance. Ye *et al.* [41] believed that the SIFT descriptor lacks deep semantic information, and DNN can make up for this defect. Their work focused on combining SIFT with DNN to achieve image registration between multispectral or multimodal image pairs. Moreover, Ma *et al.* [42] adopted a two-step strategy to achieve image registration with high accuracy. Specifically, they first regressed to coarse TPs by using a DNN and made a rough alignment for image pairs. Then, another DNN is applied to generate deep feature maps from the coarsely aligned image pairs, followed by performing image matching on the deep feature map. Recently, Zhou *et al.* [43] put multiorientated gradient features of image pairs into a pseudo-Siamese neural network in a multiscale way, which produces an auto-learned feature map for a fast and robust template matching of optical-SAR images.

2) Supervised End-to-End Methods: For supervised end-to-end methods, the loss function is usually designed based on minimizing the discrepancy between the predicted and ground truth TPs in the process of backpropagation. DeTone *et al.* [26] proposed a supervised end-to-end framework named the deep image homography estimation network (DHN), which based on Visual Geometry Group (VGG) networks to regress homography TPs for image registration. By stacking three DHN models in a cascading way, Le *et al.* [44] proposed a multiscale framework named the multiscale deep image homography estimation network (MHN), which achieves better registration performance compared with DHN. In the datasets of DHN and MHN, only single-modal image pairs with similar intensity are included. Recently, Zhao *et al.* [45] designed a loss function based on constructed feature maps named the deep Lucas–Kanade feature map (DLKFM), which makes two feature maps from multimodal images that satisfy intensity consistency. However, a practical problem for supervised methods in RSIR represents that the labels of ground truth TPs are costly and hard to obtain for real ground scenes and targets.

3) Unsupervised End-to-End Methods: For unsupervised end-to-end methods, ground truth TPs are unnecessary. The loss function of which is often designed based on optimizing the similarity between images in the process of backpropagation. These types of methods have been well-developed in medical image registration. To evaluate DVF between fixed and moving pairs of 3-D medical images, Balakrishnan *et al.* [28] proposed an unsupervised end-to-end method named VoxelMorph based on DNN. The original VoxelMorph requires that the input image pairs have been affine-registered. That is, there is no significant geometric distortion between these image pairs, only local nonuniform translations. To align images with geometric changes, Vos *et al.* [27] introduced a deep learning image registration (DLIR) framework, which inserted a network for predicting affine TPs before the networks of predicting DVF. Then, several networks for predicting DVF are stacked in a cascading way, the latter network prediction is the residual part of the previous one, and the final output DVF is obtained by integrating the predictions of all the networks. Such operation improves the registration accuracy. Furthermore, Jiang *et al.* [46] introduced multiscale fashion into the network stacking. Specifically, they stacked several networks for predicting DVF, performed downsampling of multiple scales for the input images on each network, and trained all the networks in a joint way. Finally, a coarse-to-fine registration pipeline was formed. Recently, unsupervised methods have also begun to be used in TP estimation in the field of computer vision image processing or RSIR. Nguyen *et al.* [32] propose an unsupervised learning algorithm that trains a DNN to estimate homography TPs. Huang *et al.* [47] proposed an unsupervised registration method for target tracking in SAR videos. Papadomanolaki *et al.* [48] proposed an unsupervised multistep deformable registration for aligning satellite image pairs of a single modality. Gradually, unsupervised methods have the potential for superior performance compared to the supervised methods. However, the main registration difficulty of the unsupervised methods is a large degree of geometric transformation and radiometric differences between multimodal image pairs.

III. Methodology

In this section, the proposed MU-Net is elaborated, which passes the images through several designed DNN architectures in each scale to predict the TPs and then corrects the sensed image to align with the reference one. Since the TPs are directly optimized by evaluating the similarity of structural feature descriptors of the two images, MU-Net is completely unsupervised. That means, it does not need true labels as training samples, which is different from past methods basis on supervised DL. In this article, we choose affine TPs as the form of the predicted mapping, and MU-Net can integrate other forms (e.g., homography, polynomial, and DVF) of TPs. The details are given as follows.

A. Problem Formulation

Assuming there is a pair of images f and m to be aligned. One is a reference image f with correct geographic coordinates for each pixel, and the other is a sensed image m with

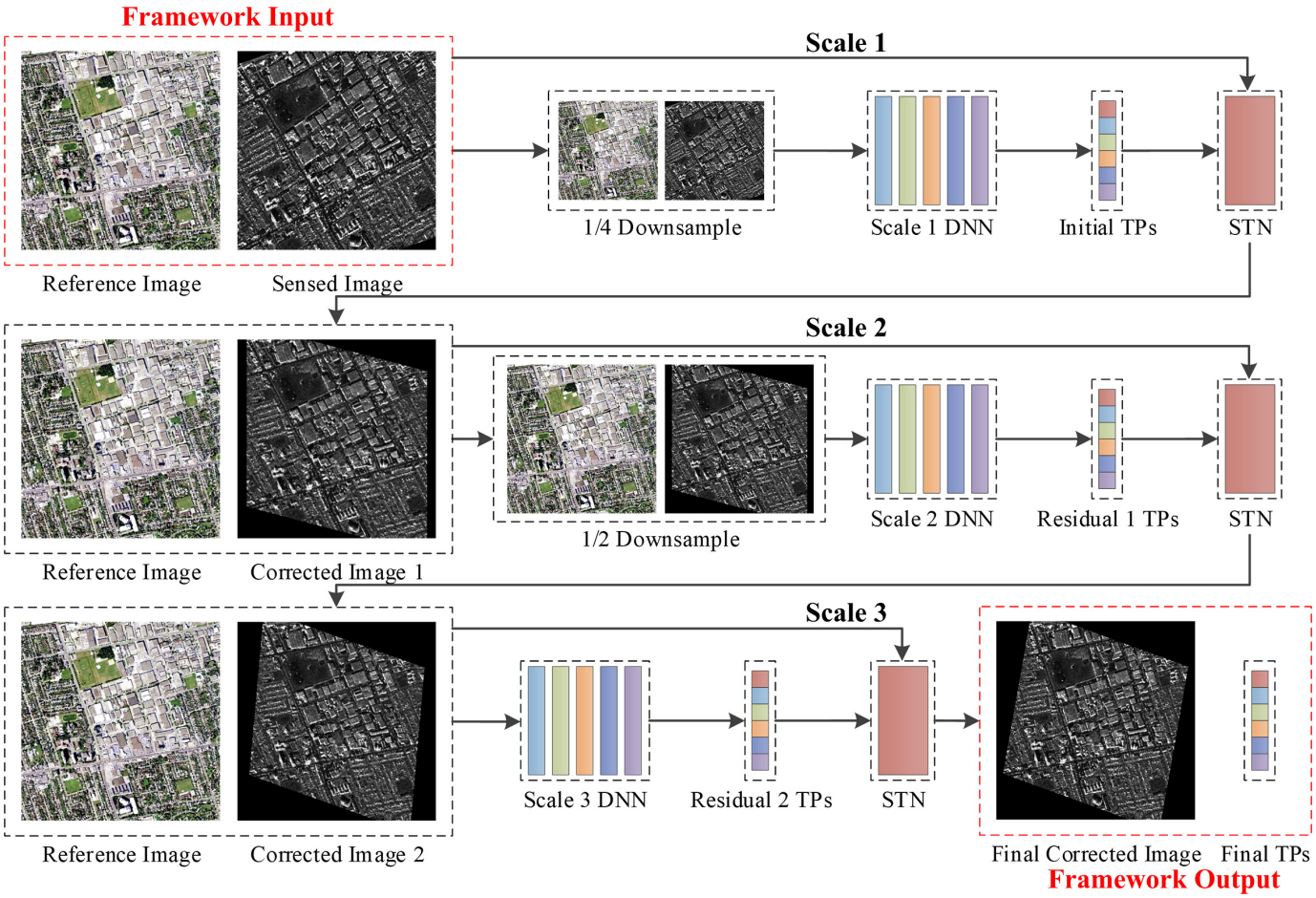

Fig. 2. Architecture of MU-Net.

geometric distortions. To correct m, the aim is to find a group of TPs μ.

In traditional image similarity optimization, μ is directly optimized by maximizing a certain similarity metric Sim

$$\hat{\mu} = \arg\max_{\mu} \left[\mathrm{Sim}\left(f, T_{\mu}(m)\right)\right] \tag{1}$$

where T_{μ} means a coordinate or spatial mapping that is parameterized by μ, and $\hat{\mu}$ is the optimal value of μ.

In the unsupervised registration method, μ is outputted by the designed DNN F

$$\mu = F_{\theta}(f, m) \tag{2}$$

where θ refers to the weights and bias parameters of F.

Therefore, μ is optimized indirectly since its θ to be directly optimized by maximizing the Sim

$$\hat{\theta} = \arg\max_{\theta} \left[\mathrm{Sim}\left(f, T_{\mu}(m)\right)\right] \tag{3}$$

where $\hat{\theta}$ is the optimal value of θ.

In MU-Net, F is defined as a stacked coarse-to-fine registration pipeline, and its weights and bias parameters θ are optimized in the process of backpropagation.

B. Multiscale Framework

In this section, we introduce the overall workflow of MU-Net. Similar to the pyramid architecture of traditional image registration methods, MU-Net performs a coarse-to-fine strategy. Three DNN models are stacked in a cascading way, and image pairs of multiple scales (e.g., different downsampling rate) are input to MU-Net to directly regress TPs, as shown in Fig. 2.

First, the DNN model in scale 1 aims to perform an initial and global alignment between the original image pairs f and m. Specifically, f and m are downsampled by a scale factor of 1/4 and inputted into the first DNN model to evaluate the initial TPs μ_1, which is applied to correct the original sensed image m to produce the first corrected sensed image $T_{\mu 1}(m)$.

Second, the DNN model in scale 2 aims to perform a residual and detailed alignment between f and $T_{\mu 1}(m)$. Specifically, f and $T_{\mu 1}(m)$ are downsampled by a scale factor of 1/2 and inputted to the second DNN model to evaluate the residual TPs $\Delta\mu_1$, which is integrated to μ_1 to yield the second TPs μ_2. And μ_2 is applied to correct the original sensed image m to produce the second corrected sensed image $T_{\mu 2}(m)$.

Third, the DNN model in scale 3 also aims to perform a further residual and detailed alignment between f and $T_{\mu 2}(m)$. Specifically, f and $T_{\mu 2}(m)$ are directly inputted to the third DNN model to evaluate the residual TPs $\Delta\mu_2$ once again, which is integrated to μ_2 to yield the final TPs μ_3. And μ_3 is applied to correct the original sensed image m to produce

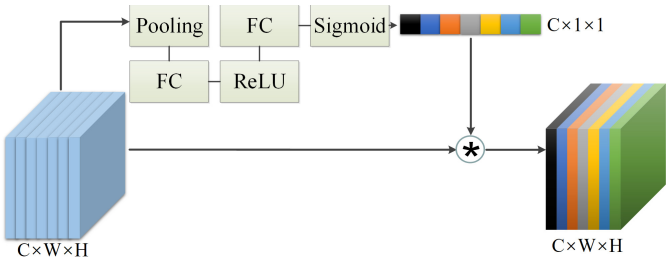

Fig. 3. Architecture of an SE block.

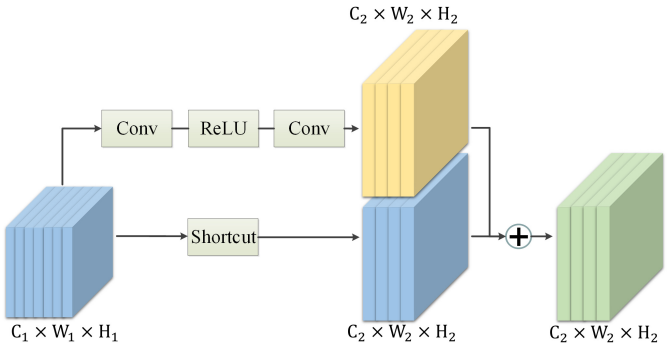

Fig. 4. Architecture of a DR block.

the final corrected sensed image $T_{\mu 3}(m)$, thus achieving the image registration.

C. Architecture of the DNN Models

In this section, we describe the designed DNN architecture for each scale adopted in our experiments, which analyzes a pair of input images globally. Note that MU-Net is universal for the DNN architecture, which has a wide range of forms and other forms may work as well as ours.

In order to extract the desired deep and semantic information from the original image and directly find the end-to-end affine mapping, we utilize the channel attention mechanism [49] and the DR network [50] in the DNN architecture. The former models the dependence between image channels and can adaptively adjust the characteristic response values of each channel. And the latter ensures that in the process of forward propagation, the deep semantic information contained in the feature map will not decrease with the deepening of the neural network layer.

1) Squeeze-and-Excitation (SE) Blocks: The process of an ordinary convolutional layer ignores the mutual relationship and interactivity among channels. The channel attention mechanism aims to facilitate the network to selectively retain features that contains a great quantity of information, so that the subsequent steps can take the advantage of these features and suppress the useless features. The SE block is one of the specific implementation forms of the channel attention mechanism. It improves the representation of a DNN model by constructing the interactivity and dependence of convolutional features in the channel direction, and allows the network to perform weight redistribution among channels.

An architecture of an SE block is shown in Fig. 3. Each SE block contains a global average pooling operation, two fully connected (FC) layers, and a simple channel-by-channel scaling operation at the end. It first examines the signal of each channel of the input feature x, compresses the global spatial information into the channel descriptor, and uses global average pooling to generate statistics z for each channel, which can be expressed as

$$z = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} x(i, j) \tag{4}$$

where (i, j) is the position of each position in the feature x on the 2-D plane, and H and W denote the height and width of x, respectively.

Second, the threshold mechanism of the sigmoid activation function is used to examine the degree of dependence of each channel. In order to limit the model complexity and enhance the generalization ability, two FC layers are used in the threshold mechanism and connected through a rectified linear unit (ReLU) activation layer δ. The first FC layer W_1 reduces the dimensionality of the features, and the number of output channels of the second FC layer W_2 is consistent with the original number C of input channels. The output of the final sigmoid function σ is the weight of each channel. This process can be expressed as

$$s = \sigma(W_2 \delta(W_1 z)) \tag{5}$$

where $W_1 \in R^{(C/r)\times C}$, $W_2 \in R^{C\times(C/r)}$, and r is a hyperparameter to reduce the number of channels.

Finally, redistributing the weight of each channel feature according to the input data itself conduces to enhance the representation of the feature

$$\hat{x} = F_{\text{scale}}(s, x) \tag{6}$$

where $F_{\text{scale}}()$ is the product operation in the channel direction and $\hat{x}$ is the final output of an SE block.

2) DR Blocks: To predict the TPs mapping directly from input image pairs, we hope that the designed DNN model is a process of information extraction. With the deepening of the network, the extracted features contain more and more deep semantic information. While the deepening of the network will also bring many defects: the calculation slows down, the model parameters are over-fitted, and the gradient disappears or explodes. The residual network solves these problems. It not only preserves the depth of the network, but also avoids the degradation problem of the deep network, that is, as the number of network layers increases, the accuracy rate is saturated or even dropped.

An architecture of a DR block is shown in Fig. 4. A DR block can be expressed as

$$x_{l+1} = H(x_l) + F(x_l, W_l) \tag{7}$$

where the DR block is divided into two parts, the direct mapping part and the residual part. $H(x_l)$ is a direct mapping, which is reflected in the curve on the bottom in Fig. 4. $F(x_l, W_l)$ is the residual part, which is generally composed of two or three convolution operations, that is, the top part of Fig. 4 contains the convolution part.

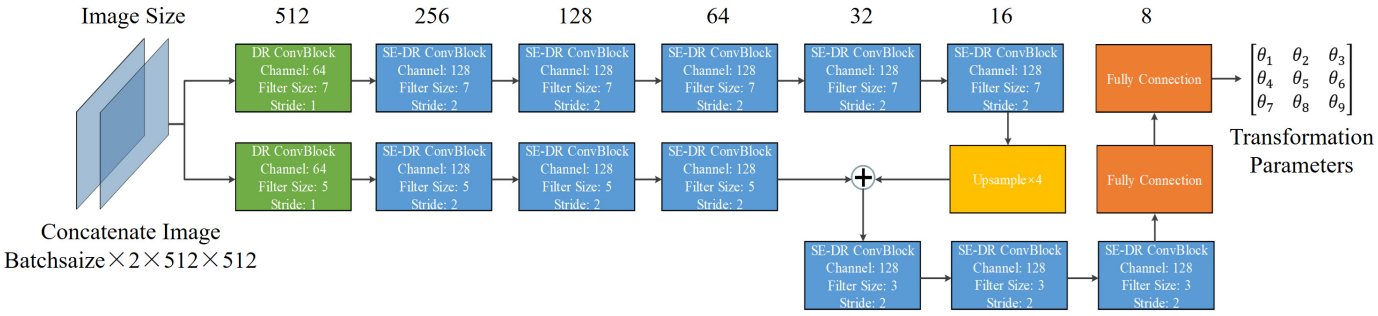

Fig. 5. Architecture of DNN architecture in scale 3.

3) Our DNN Architectures: In our experiment, an ordinary convolution block (ConvBlock) includes a convolution layer, a batch normalization layer, and a linear activation layer (Sigmoid or leaky ReLU). An SE ConvBlock is defined as an ordinary ConvBlock plus an SE block structure. And a DR ConvBlock is defined as the first convolution of the residual part of the DR block to change the width, height, and channel of the input matrix, and the subsequent convolutions keep the output size consistent with the input one. Based on the above definition, we add an SE block after each ordinary convolution in the DR ConvBlock to form an SE-DR ConvBlock, which will be used extensively in our DNN architectures. In Section IV, the ablation experiments of these ConvBlocks will be introduced.

Fig. 5 depicts the DNN architecture on the third scale. The sensed image should have the same size with the reference one, if not, zero-padding or cropping is generally adopted. Two images are concatenated in the channel direction and then passed through a series of 7×7 ConvBlocks and a series of 5×5 ones, respectively. Because when the number of channels is less than the hyperparameter r, the SE block cannot be used. Except that the first ConvBlock is a DR ConvBlock, the remaining ConvBlocks are all SE–DR ones. The two routes are concatenated by upsampling and stride connection, followed by several 3×3 SE–DR ConvBlocks. In this process, the image size is reduced while the channel of which will continue to deepen, which is conducive to extract the deep semantics information. After through the last ConvBlock, the deep semantics information is directly mapped to the affine TPs through two fully connection layers.

The DNN architectures on the first and second scales have similar structures to the DNN architecture on the third scale described above in detail. The difference is that, as we have introduced in Section III-B, the input image has undergone downsampling, so the initial size of the image has become 128×128 pixels or 256×256 pixels instead of 512×512 pixels. Therefore, we need to reduce two SE–DR ConvBlocks with a stride of 2 for each route, while maintaining the maximum number of channels at 32, which forms the DNN architecture on the first scale. Similarly, we provide for each route reduces an SE–DR ConvBlock with a stride of 2 and maintains the maximum number of channels at 64, which forms the DNN architecture on the second scale.

D. Unsupervised Training for MU-Net

In this section, we first introduce the unsupervised training implements for MU-Net, which is a joint training for DNN models of three scales. Since the quality of the DL model is highly dependent on the convergence of the loss function, we will explain in which direction the model parameters of unsupervised training are optimized. Moreover, our loss function can integrate various similarity metrics, and we adopt structural feature descriptors so that the loss function can optimize the similarity of multimodal images.

1) Training Implements and Loss Function: In MU-Net, three DNN models are stacked in a cascading way to form a combined registration pipeline. Therefore, the training procedure includes two parts, initialization and joint training.

In the first stage, to initialize the network weights of three DNN models, models on multiple scales were individually and successively trained to minimize the corresponding loss based on image structural similarity: $\text{Loss}_{\text{sim}}(f, m, \mu_1)$, $\text{Loss}_{\text{sim}}(f, m, \mu_2)$, and $\text{Loss}_{\text{sim}}(f, m, \mu_3)$. The model on the first scale was trained to regress initial TPs to achieve a rough alignment. With weights fixed for the first model, the model on the second scale was successively trained to regress residual TPs to fine-tune the alignment. Finally, the model on the third scale was trained to further correct the alignment, while freezing the weights of the first model and the second one.

In the second stage, all the weights in the stacked DNN models are unfrozen to be updatable. And each DNN model in the multiscale framework is jointly trained to collaboratively minimize the overall loss at multiple scales, which is defined as

$$\begin{aligned}\text{Loss} = {} & \lambda_1 \text{Loss}_{\text{sim}}(f, m, \mu_1) \\ & + \lambda_2 \text{Loss}_{\text{sim}}(f, m, \mu_2) + \lambda_3 \text{Loss}_{\text{sim}}(f, m, \mu_3) \quad (8)\end{aligned}$$

where λ_1, λ_2, and λ_3 are weighting factors of the loss function on multiple scales.

The consequent for image registration is that the reference image is supported to have the best similarity with the sensed image corrected by the TPs and its spatial transformation. To improve the reliability of the TPs μ, we invert the matrix of the coordinate mapping T_μ

$$T_\mu^{-1} = \left(T_\mu^T T_\mu\right)^{-1} T_\mu^T \quad (9)$$

where T_μ^{-1} denotes the inversed matrix of the coordinate mapping. Similarly, the sensed image is supported to have the

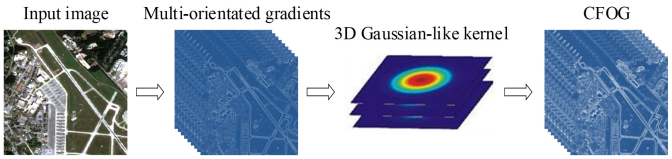

Fig. 6. Construction process of CFOG.

best similarity with the reference image warped by the inverse spatial transformation. Therefore, the similarity loss function is defined as

$$\text{Loss}_{\text{sim}}(f, m, \mu) = e^{-\left[\text{Sim}\left(f, T_{\mu}(m)\right)+\text{Sim}\left(T_{\mu}^{-1}(f), m\right)\right]/2} \quad (10)$$

where e is a constant, and Sim is the calculation method of a certain similarity metric. A higher Sim denotes a better similarity and a lower Loss_{sim}.

2) Similarity Evaluation for Multimodal Images: For multimodal RSIR, such as optical-SAR images, their pixel intensity cannot be directly used for similarity evaluation due to nonlinear radiometric differences. Considering that structure features are preserved between multimodal images [35], [39], [51], we use the structural descriptor instead of intensity to calculate the value of similarity metric between multimodal images.

In our experiment, we mainly adopted the similarity metric and the structural descriptor chanel feature of orientated gradient (CFOG) [14] to convergence the loss function, and note that MU-Net can also integrate other feature descriptors as working well. CFOG is a fast and robust structural descriptor with reliable matching performance and high computational efficiency, which first extracts multioriented gradients and then constructs the oriented histogram. On the basis of the oriented histogram, the convolution operation is performed by a 3-D Gaussian-like kernel, which collects the orientated gradients of neighboring pixels. Thus, a 3-D structural feature map is generated. Fig. 6 depicts this construction process.

We adopt NCC for $\text{Sim}(A, B)$ on the structural feature maps A and B. NCC determines the correspondences between two feature maps by searching the location of the maximum value. In general, the NCC can be computed as

$$\text{NCC}(A, B) = \frac{\sum_{p\in N}\left(A(p)-\tilde{A}\right)\left(B(p)-\tilde{B}\right)}{\sqrt{\sum_{p\in N}\left(A(p)-\tilde{A}\right)}\sqrt{\sum_{p\in N}\left(B(p)-\tilde{B}\right)}} \quad (11)$$

where $\tilde{A}$ and $\tilde{B}$ denotes the mean intensity of the reference and the sensed feature maps, respectively. The value of NCC is in the interval $[-1, 1]$, and a higher value of NCC denotes a higher similarity.

IV. Experiment

A. Datasets and Comparison Methods

1) Generation of Image Pairs: To evaluate MU-Net by a large number of experimental data, we introduce the generation process of image pairs, which are used to augment the training and testing data in our experiment. As an example shown in Fig. 7, I_1 and I_2 are two precisely aligned images with a size of larger than 512×512 pixels. I_2 is randomly

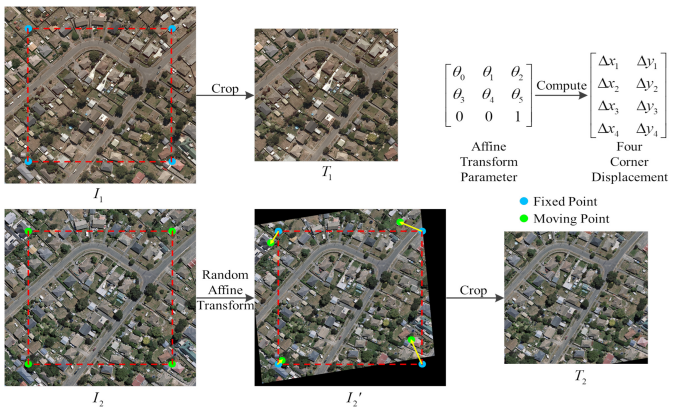

Fig. 7. Example of generation of reference and sensed image pairs.

performed affine transformation to generate warped image I_2'. Then, the same positions of I_1 and I_2' are cropped, which obtain the reference image T_1 and the sensed image T_2 as the image pair with the same size of 512×512 pixels. The ground truth corner displacement $[\Delta x_i, \Delta y_i]$ between the four corner points in I_1 and their corresponding position in I_2' can be calculated by the affine TPs

$$\begin{bmatrix}\Delta x_i \\ \Delta y_i \\ 1\end{bmatrix} = \left(\begin{bmatrix}\theta_0 & \theta_1 & \theta_2 \\ \theta_3 & \theta_4 & \theta_5 \\ 0 & 0 & 1\end{bmatrix} - E\right)\begin{bmatrix}x_i \\ y_i \\ 1\end{bmatrix} \quad (12)$$

where $[x_i, y_i]$ represents the position of the ith corner point, and E represents the identity matrix. The four ground truth corner displacements are used to subsequently evaluate the registration accuracy. A large number of image pairs are generated in this way, which can increase the datasets for our subsequent experiments. The specific experimental datasets will be introduced in detail as follows.

2) Datasets: MU-Net has been extensively evaluated on several types of multitemporal or multimodal remote sensing datasets, including optical-optical datasets, optical-infrared datasets, optical-SAR datasets (including medium- and high-resolution data), and optical-map datasets. These datasets are composed of a large number of image pairs generated according to the generation process mentioned above. These datasets are used for two purposes. First, we want to show that MU-Net is generally applicable to registration tasks for various types of multimodal remote sensing images. In addition, the registration difficulty of the five multimodal remote sensing images is gradually increasing. In particular, since SAR images contain obvious speckle noise and raster maps contain text labels, the registration of optical-SAR and optical-map pairs is more challenging than other cases.

a) Optical-optical: WHU building dataset is a set of multitemporal aerial photographs, which consists of two pre-registered aerial images and the change masks captured on the same area of Christchurch, New Zealand, in 2012 and 2016, respectively. The size of each image is $32\,507 \times 15\,345$ pixels with a resolution of 0.075 m. The ground object, colors, and textures on two images are different with the changes in time and aerial photography conditions. Since many areas

with building have changed in the two images, we manually selected the areas with only a small amount of changes in ground objects. We crop these areas into about 500 image pairs with a size of 600 × 600 pixels. For each pair of images, we perform ten random affine transformations and center crops on them. The final optical-optical dataset includes about 5000 pairs of multitemporal images with a size of 512 × 512 pixels. Approximately, 4500 optical-optical image pairs are used for training and 500 for testing.

b) Optical-infrared: The image pairs in optical-infrared dataset are derived from several Landsat-8 satellite images with a size of approximately 7600 × 7800 pixels. The coverage of which includes the Chengdu Plain and surrounding hills and mountains with a resolution of 30 m. The optical images were acquired by Landsat-8 band 2 in July 2020, and the corresponding infrared images were acquired by Landsat-8 band 5 in February 2021. The image has undergone geographic correction and manual geometric correction to ensure affine alignment. We cropped the large size image into 500 small image pairs with a size of 700 × 700 pixels. For each pair of images, we perform ten random affine transformations and center crops on them. The final optical-infrared dataset includes about 5000 pairs of multimodal images with a size of 512 × 512 pixels. Approximately, 4500 optical-infrared image pairs are used for training and 500 for testing.

c) Optical-SAR (high resolution): This optical-SAR dataset (high resolution) are derived from coregistered image pairs with a resolution of 3 m, and the coverage of which includes Chengdu and Ningbo in China. The SAR images and the optical images are acquired by GF3 and ZY3, respectively. Similar to the above operation, 1840 image pairs are used for training and 460 for testing.

d) Optical-SAR (medium resolution): The coverage of optical-SAR dataset (medium resolution) includes difference scenes such as cities, farmland, rivers, and forests. The image pair is acquired by Sentinel-1 and Sentinel-2 in May 2021 with a resolution of 10 m. Around 5000 image pairs are used for training and 800 for testing.

e) Optical-map: We obtained the optical images and the corresponding Google maps with a resolution of 1 m from the Google map service. The area where the image is located is in Tokyo, and the features contained are mainly dense buildings and streets. The maps of these scenes and the corresponding optical images have similar structure. Around 5000 optical-map pairs are used for training and 800 for testing.

3) Comparison Methods: As the registration difficulty of the several modalities of image pairs is gradually increasing, we could observe the trend of changes in the performance of MU-Net compared with the state-of-the-arts methods (i.e., SIFT, RIFT, CFOG, DLKFM, and DLIR), and evaluate their flexibility for different types of multimodal image registration. Table I briefly introduces and classifies the mentioned methods.

TABLE I
BRIEF INTRODUCTION AND CLASSIFICATION FOR THE METHODS OF COMPARISON

Method	Type	Classification
SIFT	Traditional (Handicraft)	Feature Matching
RIFT	Traditional (Handicraft)	Feature Matching
CFOG	Traditional (Handicraft)	Template Matching
DLKFM	Deep Learning	Supervised End-to-End Mapping
DLIR	Deep Learning	Unsupervised End-to-End Mapping
MU-Net	Deep Learning	Unsupervised End-to-End Mapping

B. Evaluation Criteria and Implementation Details

1) Evaluation Criteria: Between the ground truth affine TPs and the predicted ones, it is difficult to digitally balance the translation, rotation, scaling, or shearing components by directly calculating the differences. Therefore, (12) is adopted to transform the affine TPs into the four corner displacements and then calculate the differences between the ground truth four corner displacements and the predicted ones. Similar to the recent literatures of end-to-end learning for image registration [26], [44], [45], we use the average corner error (ACE) as the evaluation criteria of the registration accuracy, which is defined as the root-mean-square error (RMSE) between the ground truth four corner displacements $[\Delta x_i^{\text{gt}}, \Delta y_i^{\text{gt}}]$ and the predicted ones $[\Delta x_i^{\text{pre}}, \Delta y_i^{\text{pre}}]$

$$e_c = \frac{1}{4}\sum_{i=1}^{4}\sqrt{\left(\Delta x_i^{\text{gt}} - \Delta x_i^{\text{pre}}\right)^2 + \left(\Delta \mathrm{y}_i^{\text{gt}} - \Delta y_i^{\text{pre}}\right)^2}. \quad (13)$$

In addition, when applying random affine transformation to the original images and generating the augments samples, we can obtain the ground truth affine TPs, which are used to calculate the RMSE. Note that these ground truth TPs are just used for accuracy evaluation and are not utilized for the training of MU-Net.

2) Implementation Details: The experimental parameters used the following settings. The NCC on the CFOG feature maps is adopted as the similarity metric for MU-Net. Some random affine transformations consist of random rotation, translation, scale, and shear transformations, where the scale parameter is limited in a range of [0.5, 2] with a precision of 0.1, the translation parameter is limited in a range of [−0.1, 0.1] with a precision of 0.002 (about 1 pixel in an image with a size of 512 × 512 pixels), the rotation parameter is limited in a range of $[-\pi, \pi]$ with a precision of 1°, and the shear parameter is limited in a range of $[-\pi/6, \pi/6]$ with a precision of 1°. The training takes 500 iterations, the initial learning rate is set as 0.002, and the weight decay is set as 0.005. The weighting factors of the loss function at multiscale levels are set to [0.05, 0.05, 0.9].

For compared methods, we use their parameter settings recommended by the authors. In our experiments, the traditional matching methods (i.e., SIFT, RIFT, and CFOG) adopt RANSAC to fit the affine TPs. The affine transformation is a form of TPs, which is commonly used in RSIR task. In MU-Net, all DNN architectures on each scale predict six

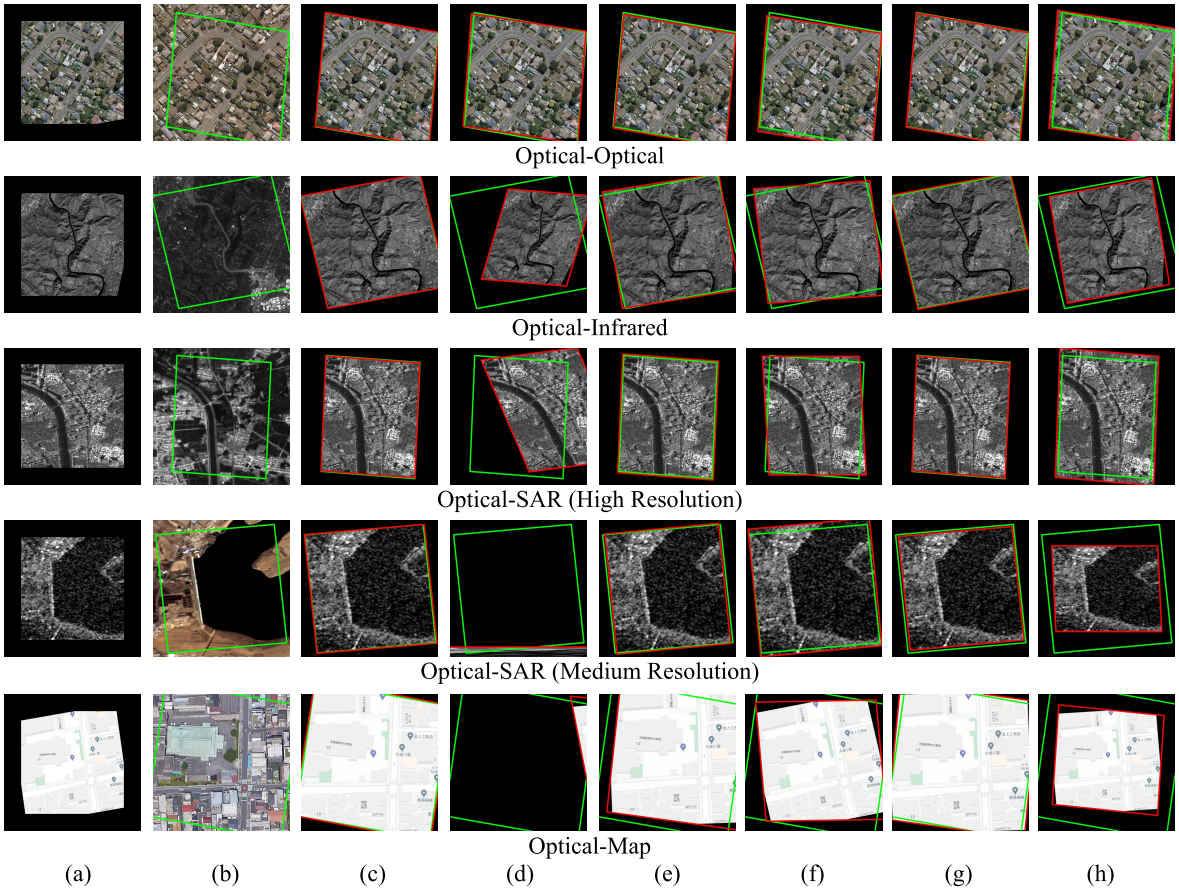

Fig. 8. (From top to bottom) Registration results of an example on the optical-optical, optical-infrared, optical-SAR (high resolution), optical-SAR (medium resolution), and optical-map datasets, respectively: (a) randomly affine warped and cropped sensed image; (b) reference image; (c) MU-Net; (d) SIFT; (e) RIFT; (f) CFOG; (g) DLKFM; and (h) DLIR. The green line represents the ground truth registration result, and the red line represents the experimental registration result for each method.

affine TPs, which include the translation, rotation, scaling, and shearing deformation.

All resampling operations in the experiment were performed by bilinear interpolation and implemented on STN. And all experiments were performed in Pytorch 1.8.0 [52] basis on Python 3.8 on an NVIDIA GeForce RTX 3080 GPU, an Intel Core i7-10700KF 3.80-GHz CPU with 8 cores (16 threads).

C. Accuracy Analysis

In this section, we use the test datasets introduced in Section IV-A to compare MU-Net with the other methods. Between the image pair of these datasets, the sensed image has a random affine transformation distortion relative to the reference image. Here, we separately conduct qualitative and quantitative analyses for the above methods on the five datasets.

1) Qualitative Accuracy Analysis: Fig. 8 shows five registration examples on five multitemporal or multimodal datasets. The red lines indicate the experimental test results and the green ones indicate the ground truth registration, respectively. In general, MU-Net shows that the red lines mostly overlap with the green lines, which illustrates the best registration accuracy. From the generalization performance of the method, all methods achieve image rough alignment on the optical-optical dataset. With the increasing difficulty of registration on multimodal datasets, some comparison methods fail. For example, SIFT and DLIR cannot match some multimodal images at all. CFOG also has many failures, which are caused by significant geometric differences between images. It is observed that RIFT achieves some robustness to geometric distortions and radiometric differences on the first two datasets, while still greatly decease its accuracy on the latter three datasets. As a supervised learning method, DLKFM shows comparable results that the red lines and green ones almost overlap. However, the strength of MU-Net is that it is data-driven and completely unsupervised.

2) Quantitative Accuracy Analysis: For each method tested on different datasets, Table III lists the percentage of the test images with the ACE smaller than 3, 5, 10, and 20 pixels. On five datasets, MU-Net achieves the best registration accuracy. However, it is incomplete to compare the above methods by directly using image pairs that have translation, scale, and rotation distortions at the same time. Subsequently, we carried out additional experiments to gradually increase the distortion on the sensed image in the three types of translation, rotation, and scale, and compare the flexibility against these distortions among MU-Net and the other methods. Figs. 9–11 show the ACE of each method for translation, rotation, and scale deformation, respectively. Similarly, these experiments are performed on different multimodal image datasets. We can observe and evaluate the registration performance from simple

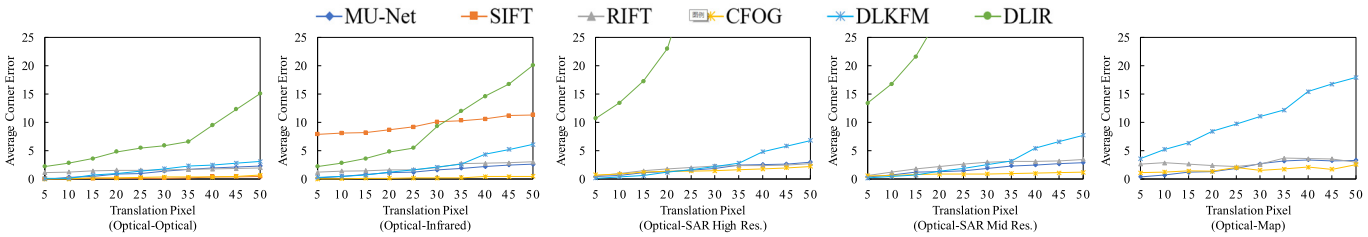

Fig. 9. Comparison of ACE for translation distortion on five datasets.

TABLE II
COMPARISON OF EVALUATION FOR MENTIONED METHODS

Dataset	Method	The percentage of ACE (pixels)			
		≤3	≤5	≤10	≤20
Optical-Optical	MU-Net	96%	97%	98%	98%
	SIFT	51%	69%	88%	90%
	RIFT	63%	78%	91%	91%
	CFOG	15%	19%	25%	52%
	DLKFM	92%	96%	98%	98%
	DLIR	82%	86%	97%	98%
Optical-Infrared	MU-Net	96%	97%	98%	98%
	SIFT	39%	51%	57%	59%
	RIFT	54%	67%	88%	89%
	CFOG	14%	17%	23%	51%
	DLKFM	63%	84%	91%	98%
	DLIR	1%	8%	72%	97%
Optical-SAR (High Res.)	MU-Net	92%	95%	95%	97%
	SIFT	0%	0%	0%	0%
	RIFT	57%	61%	84%	88%
	CFOG	15%	21%	27%	55%
	DLKFM	67%	83%	93%	97%
	DLIR	0%	4%	32%	53%
Optical-SAR (Med. Res.)	MU-Net	92%	94%	95%	97%
	SIFT	0%	0%	0%	0%
	RIFT	53%	68%	90%	92%
	CFOG	14%	17%	25%	51%
	DLKFM	59%	79%	90%	97%
	DLIR	0%	2%	23%	67%
Optical-Map	MU-Net	89%	92%	94%	96%
	SIFT	0%	0%	0%	0%
	RIFT	17%	34%	57%	61%
	CFOG	12%	15%	25%	49%
	DLKFM	56%	74%	89%	97%
	DLIR	0%	0%	21%	29%

to difficult data. Accordingly, further discussions about each method are given as follows.

SIFT This is one of the classic feature matching methods. From Table II, we can see that SIFT has a certain accuracy on the optical-optical dataset, but its performance on the optical-infrared dataset is greatly declined. In addition, SIFT completely fails on the optical-SAR and the optical-map datasets. From Figs. 9–11, we can observe that SIFT is robust to translation, scale, and rotation differences between single-modal images (e.g., optical-optical), but it is vulnerable to multimodal image matching (e.g., optical-SAR).

RIFT: This is a feature matching method for multimodal remote sensing images. It can be seen from Table II that there is no significant change in the registration performance for different types of multimodal datasets. As can be seen from Figs. 9–11, RIFT has good translation and rotation invariance, but it cannot handle images with scale changes.

CFOG: This is a template matching method with fast computational efficiency and robust matching performance. From Fig. 9, we can see that in the case of only translation distortions, the registration performance of CFOG is the best, and it is robust to radiometric differences for multimodal images with inconsistent translations, whereas Figs. 10 and 11 show that CFOG is sensitive to rotation and scale differences.

DLKFM: This method adopts supervised end-to-end learning, which trains its network with the L2 distance between the predicted TPs and the ground truth TPs as the loss function. From Table II, DLKFM aligns 92% of test image pairs on optical-optical images within a three-pixel ACE. Nevertheless, its performance on the three other multimodal datasets has dropped significantly. On these datasets, the percentage of ACE within three pixels is drastically reduced, thus indicating that the radiometric resistance learned by DLKFM has a certain limitation. From Figs. 9–11, DLKFM is robust to translation and rotation differences but easily influenced by radiometric changes.

DLIR: This is a pioneering method of unsupervised learning in the field of medical imaging. Since this method is to predict DVF for dense matching, we adopted the framework and network model introduced in the original author's article, but in the end, we returned the affine TPs instead of DVF. From Table II, DLIR aligns 82% of test image pairs on optical-optical datasets within a three-pixel ACE, but it achieves poor performance for the other four multimodal datasets. According to [27], DLIR is suitable for image pairs with roughly affine alignment. From Figs. 9–11, we can see that DLIR is not suitable for the image registration with significant geometric distortions, even the simplest distorted image cannot be corrected accurately.

MU-Net: From Table II, MU-Net obtains the best accuracy, and the accuracy is almost not affected by different image modalities. For example, on the optical-optical dataset that is the easiest to be aligned, MU-Net aligns 96% of test

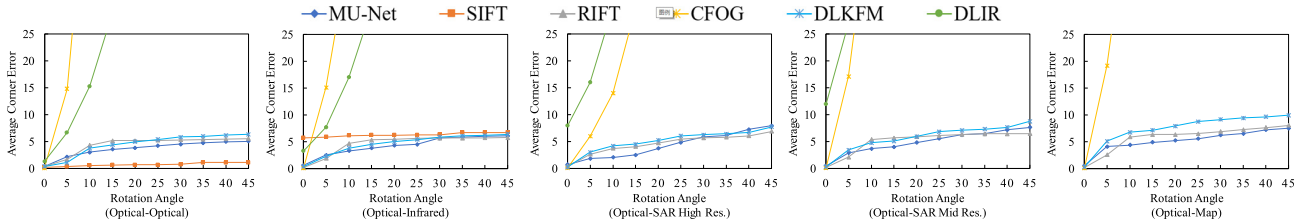

Fig. 10. Comparison of ACE for rotation distortion on five datasets.

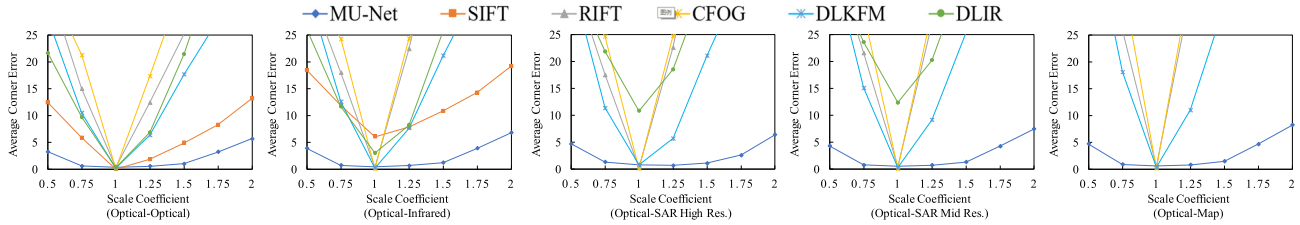

Fig. 11. Comparison of ACE for scale distortion on five datasets.

image pairs within a three-pixel ACE and 89% ones even on the optical-map dataset that is the most difficult for image registration. From Fig. 9, on the test images of all modalities, the accuracy of MU-Net within the translation difference of 30 pixels is equivalent to that of CFOG. When the translation differences are larger than 30 pixels, the performance is slightly degraded. A general disadvantage of end-to-end learning is that the larger the initial distortion is, the larger the error of the prediction result is. Nevertheless, MU-Net is still superior to the other compared methods. From Fig. 10, on the optical-optical and optical-infrared datasets, our method is comparable to the supervised DLKFM and outperforms the RIFT. As the registration difficulty increases, our method is less affected by rotation changes on the optical-SAR and optical-map datasets and presents the best registration performance. It can be clearly observed from Fig. 11 that MU-Net obtains the best performance in image registration with scale distortions, and its performance on the optical-optical datasets is even better than SIFT, and it is also competent for the other three types of multimodal image registration tasks with scale changes. On the five datasets, the errors do not change significantly with modal changes. In general, MU-Net can solve the registration problem with a scale change in the range of [0.5, 2], and the registration error is within five pixels.

The above experimental results prove that MU-Net can be applied to various types of multimodal image registration tasks and can align image pairs with translation, rotation, scale, and radiation changes. Comparing various traditional feature-based and area-based methods, supervised learning methods, and other unsupervised learning methods, MU-Net achieves the most comprehensive and accurate registration results overall.

D. Analysis of Noise Sensitivity

This section examines the noise sensitivity of MU-Net by adding the Gaussian white noise. The percentage of the test images within a three-pixel ACE is used to analyze the noise sensitivity. For each image pairs on the optical-optical and optical-SAR test datasets (medium resolution), we add the Gaussian white noise with a mean value of 0 and a variance in the range of [0, 1%] to the sensed image to generate a series of noisy sensed images.

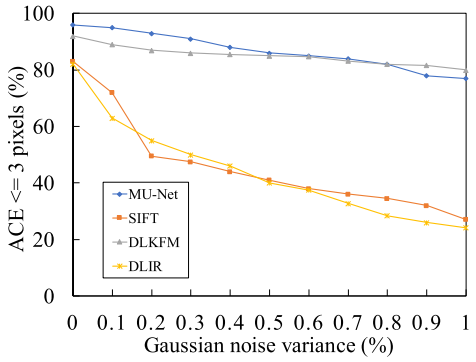

Fig. 12. Percentage of images with ACE smaller than three pixels versus various Gaussian noise on the optical-optical dataset.

On the optical-optical dataset, we compare MU-Net with SIFT, DLIR, and DLFKM. On the one hand, the experimental results in Section IV-C have shown that the above methods can handle certain affine distortions on the optical-optical image dataset. On the other hand, there is no significant radiometric difference on the optical-optical dataset, which makes that the experiment can objectively evaluate the influence of noise for these methods. Fig. 12 shows the percentage of ACE $\leq$ 3 pixels versus various Gaussian noise in image registration on the optical-optical dataset. MU-Net and DLKFM perform better than other methods under various Gaussian noise, which indicates that the ability of MU-Net to resist noise can achieve the comparable performance of the supervised method, or even better within a certain range. Although SIFT can handle affine distortions between single-modal images, it can less resist the Gaussian noise. As an unsupervised method, DLIR presents the highest noise sensitivity, whereas MU-Net overcomes this shortcoming.

On the optical-SAR dataset (medium resolution), SAR images are inherently noisy even without adding artificial noise. Therefore, we wish to show MU-Net remains robust in more difficult situations. Since other methods (e.g., SIFT, RIFT, CFOG, and DLIR) cannot achieve comparable accuracy

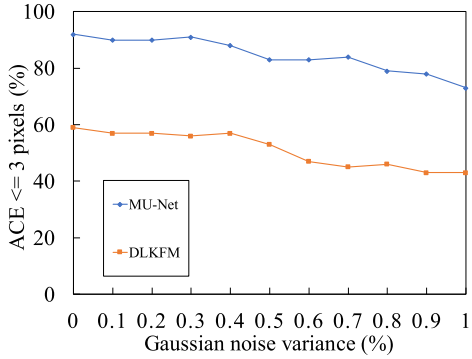

Fig. 13. Percentage of images with ACE smaller than three pixels versus various Gaussian noise on the optical-SAR dataset (medium resolution).

TABLE III
EFFECTS OF SE, DR, AND SE–DR CONVBLOCKS ON ACCURACY

ConvBlock Types	The percentage of ACE (pixels)			
	≤3	≤5	≤10	≤20
Ordinary ConvBlocks	80%	88%	93%	95%
SE ConvBlocks	88%	92%	94%	96%
DR ConvBlocks	82%	90%	93%	95%
SE-DR ConvBlocks	92%	94%	95%	97%

on the optical-SAR dataset, we compare MU-Net only with DLKFM. Fig. 13 shows the percentage of ACE $\leq$ 3 pixels versus various Gaussian noise in image registration on the optical-SAR dataset (medium resolution). In more difficult situations, the accuracy of two methods drops slowly, while our method still remains higher accuracy. We ascribe this to the very robust and consistent local descriptor on the loss function of MU-Net. In general, the experiment of noise sensitivity validates the generalized noise robustness of MU-Net.

E. Ablation Study

1) Effectiveness of SE-DR ConvBlock: In Section IV, we have introduced the specific architectures of the DNN model designed for each scale. These architectures adopt SE–DR ConvBlocks that combine the channel attention mechanism and the DR network. In order to evaluate the positive effects of SE blocks and DR blocks on the DNN model architectures, we use ordinary ConvBlocks, SE-blocks, and DR-blocks to replace SE–DR blocks in the DNN model, and evaluate the accuracy on the optical-SAR dataset (medium resolution). The detailed experimental results are shown in Table III.

From Table III, compared with the architectures that only consist of ordinary ConvBlocks, the addition of SE ConvBlocks results in a significant improvement on accuracy. Meanwhile, the adoption of DR ConvBlock also positively improves the accuracy. Therefore, considering these positive effects, SE-DR ConvBlocks are adopted in the DNN architectures of MU-Net.

2) Effectiveness of Multiscale Cascading: The number of stacked DNN models with different scales is one of the important hyperparameters of our multiscale registration framework. Adopting the same learning parameter settings, we evaluate the impact of cascading different numbers of DNN models with different scales on the registration accuracy. The experimental results on the optical-SAR dataset (medium resolution) are shown in Table IV.

TABLE IV
EFFECTS OF THE NUMBER OF CASCADING DNN MODELS ON ACCURACY WHERE THE NUMBERS IN BRACKETS INDICATE THE MULTIPLES OF DOWNSAMPLING

Number of cascading DNN models	The percentage of ACE (pixels)			
	≤3	≤5	≤10	≤20
1 scale (1)	41%	68%	89%	90%
2 scales (1/2, 1)	63%	81%	93%	95%
3 scales (1/4, 1/2, 1)	92%	94%	95%	97%
4 scales (1/8, 1/4, 1/2, 1)	77%	86%	93%	95%

TABLE V
EFFECTS OF DIFFERENT WEIGHTING FACTORS ON ACCURACY

Weighting Factors (λ_1, λ_2, λ_3)	The percentage of ACE (pixels)			
	≤3	≤5	≤10	≤20
0.2, 0.2, 0.6	85%	90%	94%	94%
0.1, 0.1, 0.8	89%	92%	95%	96%
0.05, 0.05, 0.9	92%	94%	95%	97%
0.025, 0.025, 0.95	63%	79%	91%	92%

Table IV shows that when the number of cascading DNN models increases from one to three, the registration accuracy is greatly improved. Meanwhile, MU-Net can also handle a larger range of image distortions. But when the number of models increases to four, the performance has not been further improved, where the registration accuracy is decreased. We ascribe this phenomenon to the very small size of the image input to the first DNN model, which makes it difficult to extract deep semantic information. Therefore, three DNN models with different scales are adopted in MU-Net.

3) Effectiveness of the Loss Function: The weighting factors of the loss function on multiple scales are another important hyperparameters. Adopting the same learning parameter settings, we evaluate the impact of different weighting factors on the registration accuracy. The experimental results on the optical-SAR dataset (medium resolution) are shown in Table V.

From Table V, it is observed that with the weighting factor on scale 3 increased from 0.6 to 0.9, the registration accuracy is gradually improved. Based on this, if we continue to increase the weighting factor on scale 3, the registration accuracy decreases rapidly. This ablation study illustrates that scale 3 plays the role of global alignment for image registration, while scale 1 and scale 2 play the role of refined alignment. Therefore, the weighting factors of the loss function at multiscale levels are set to [0.05, 0.05, 0.9] in MU-Net.

V. CONCLUSION

In this article, we propose an MU-Net for RSIR. Without the ground truth labels, MU-Net directly learns the end-to-end mapping from the image pairs to their TPs. MU-Net stacks several DNN models on their scales to avoid being trapped in a local extremum and resist a large degree of image distortions (including geometry and radiation). We design a novel loss

function paradigm based on structural similarity, which makes MU-Net suitable for various types of multimodal images.

Experiments are performed on virous types of datasets, the optical-optical, optical-infrared, optical-SAR, and optical-map. Experimental results show that MU-Net has the most comprehensive and accurate performance to be robust to geometric and radiometric distortions between multimodal image pairs, compared with the current state-of-the-art traditional methods (such as SIFT, CFOG, and RIFT) and supervised and unsupervised DL methods (such as DLIR and DLKFM).

MU-Net can be regarded as a common framework for RSIR. This is because it can regress the parameters of various transformation models (e.g., homography, polynomial, and DVF) beside the used affine model. Moreover, other structure feature descriptors such as HOPC [35] and the histogram of oriented gradient (HOG) [53] can also be integrated as similarity metrics for calculating the loss function in MU-Net.

In our work, we also found some failure cases in special conditions, which may guide researchers to improve it in kinds of applications. First, if the scales of the input images are so different that the overlap area is less than about 50%, the loss function may not converge or be trapped into local optimization. Second, the texture of some images is very inconspicuous, and it is difficult to extract their structural features. In this case, the loss function will converge in the wrong position.

In addition, there are still many research directions for unsupervised end-to-end learning, such as the generalization of the network model, the estimation of the initial transformation, and the judgment conditions for stopping iteration or failure to converge. Despite this, our works have proved the potential of unsupervised end-to-end learning frameworks or methods in RSIR tasks. Therefore, further research on this topic is encouraged.

REFERENCES

[1] J. Ma, X. Jiang, A. Fan, J. Jiang, and J. Yan, "Image matching from handcrafted to deep features: A survey," *Int. J. Comput. Vis.*, vol. 129, pp. 23–79, Aug. 2020.

[2] 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, 2021.

[3] W. Ma *et al.*, "Remote sensing image registration with modified sift and enhanced feature matching," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 1, pp. 3–7, Jan. 2017.

[4] H. J. Johnson and G. E. Christensen, "Consistent landmark and intensity-based image registration," *IEEE Trans. Med. Imag.*, vol. 21, no. 5, pp. 450–461, May 2002.

[5] K. Mikolajczyk *et al.*, "A comparison of affine region detectors," *Int. J. Comput. Vis.*, vol. 65, no. 1, pp. 43–72, Nov. 2005.

[6] E. Dubrofsky, "Homography estimation," Ph.D. dissertation, Diplomová Práce, Univerzita Britské Kolumbie, Vancouver, BC, Canada, 2009. [Online]. Available: https://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.186.4411

[7] X. Cao *et al.*, "Deformable image registration based on similarity-steered CNN regression," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2017, pp. 300–308, doi: 10.1007/978-3-319-66182-7_35.

[8] M. A. Fischler and R. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," *Commun. ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[9] D. Barath, J. Matas, and J. Noskova, "MAGSAC: Marginalizing sample consensus," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 10197–10205.

[10] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.

[11] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-up robust features (SURF)," *Comput. Vis. Image Understand.*, vol. 110, no. 3, pp. 346–359, 2008.

[12] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "ORB: An efficient alternative to SIFT or SURF," in *Proc. Int. Conf. Comput. Vis.*, Nov. 2011, pp. 2564–2571.

[13] M. Gesto-Diaz *et al.*, "Feature matching evaluation for multi-modal correspondence," *ISPRS-J. Photogramm. Remote Sens.*, vol. 129, pp. 179–188, 2017.

[14] Y. Ye, L. Bruzzone, J. Shan, F. Bovolo, and Q. Zhu, "Fast and robust matching for multimodal remote sensing image registration," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 11, pp. 9059–9070, Nov. 2019.

[15] M. P. Heinrich *et al.*, "MIND: Modality independent neighbourhood descriptor for multi-modal deformable registration," *Med. Image Anal.*, vol. 16, no. 7, pp. 1423–1435, 2012.

[16] B. S. Reddy and B. N. Chatterji, "An FFT-based technique for translation, rotation, and scale-invariant image registration," *IEEE Trans. Image Process.*, vol. 5, no. 8, pp. 1266–1271, Aug. 1996.

[17] J. Li, Q. Hu, and M. Ai, "RIFT: Multi-modal image matching based on radiation-variation insensitive feature transform," *IEEE Trans. Image Process.*, vol. 29, pp. 3296–3310, 2020.

[18] 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.

[19] Y. Li, Y. Zhang, X. Huang, and J. Ma, "Learning source-invariant deep hashing convolutional neural networks for cross-source remote sensing image retrieval," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 11, pp. 6521–6536, Nov. 2018.

[20] Y. Li, J. Ma, and Y. Zhang, "Image retrieval from remote sensing big data: A survey," *Inf. Fusion*, vol. 67, pp. 94–115, Mar. 2021.

[21] S. Hao, W. Wang, Y. Ye, E. Li, and L. Bruzzone, "A deep network architecture for super-resolution-aided hyperspectral image classification with classwise loss," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 8, pp. 4650–4663, Aug. 2018.

[22] S. Hao, W. Wang, Y. Ye, T. Nie, and L. Bruzzone, "Two-stream deep architecture for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 4, pp. 2349–2361, Apr. 2018.

[23] Y. Ye *et al.*, "An adaptive attention fusion mechanism convolutional network for object detection in remote sensing images," *Remote Sens.*, vol. 14, no. 3, p. 516, Jan. 2022.

[24] X. Dong, R. Fu, Y. Gao, Y. Qin, Y. Ye, and B. Li, "Remote sensing object detection based on receptive field expansion block," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022, doi: 10.1109/LGRS.2021.3110584.

[25] Z. Yang, T. Dan, and Y. Yang, "Multi-temporal remote sensing image registration using deep convolutional features," *IEEE Access*, vol. 6, pp. 38544–38555, 2018.

[26] D. DeTone, T. Malisiewicz, and A. Rabinovich, "Deep image homography estimation," 2016, *arXiv:1606.03798*.

[27] B. D. de Vos, F. F. Berendsen, M. A. Viergever, H. Sokooti, M. Staring, and I. Išgum, "A deep learning framework for unsupervised affine and deformable image registration," *Med. Image. Anal.*, vol. 52, pp. 128–143, Feb. 2019.

[28] G. Balakrishnan, A. Zhao, M. R. Sabuncu, J. Guttag, and A. V. Dalca, "VoxelMorph: A learning framework for deformable medical image registration," *IEEE Trans. Med. Imag.*, vol. 38, no. 8, pp. 1788–1800, Aug. 2019.

[29] W. Liu, Z. Wang, X. Liu, N. Zeng, Y. Liu, and F. E. Alsaadi, "A survey of deep neural network architectures and their applications," *Neurocomputing*, vol. 234, pp. 11–26, Apr. 2017.

[30] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu, "Spatial transformer networks," 2015, *arXiv:1506.02025*.

[31] Y. Fu *et al.*, "Deep learning in medical image registration: A review," *Phys. Med. Biol.*, vol. 65, no. 20, Oct. 2020, Art. no. 20TR01.

[32] T. Nguyen, S. W. Chen, S. S. Shivakumar, C. J. Taylor, and V. Kumar, "Unsupervised deep homography: A fast and robust homography estimation model," *IEEE Robot. Autom. Lett.*, vol. 3, no. 3, pp. 2346–2353, Jul. 2018.

[33] A. V. DalcaEmail *et al.*, "Unsupervised learning for fast probabilistic diffeomorphic registration," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2018, pp. 729–738.

[34] A. Sedaghat, M. Mokhtarzade, and H. Ebadi, "Uniform robust scale-invariant feature matching for optical remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 49, no. 11, pp. 4516–4527, Nov. 2011.

[35] Y. Ye and L. Shen, "Hopc: A novel similarity metric based on geometric structural properties for multi-modal remote sensing image matching," *ISPRS Ann. Photogramm., Remote Sens. Spatial Inf. Sci.*, vol. 3, pp. 9–16, Jul. 2016.

[36] Y. Xiang, F. Wang, and H. You, "OS-SIFT: A robust SIFT-like algorithm for high-resolution optical-to-SAR image registration in suburban areas," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 8, pp. 3078–3090, Jun. 2018.
[37] J. N. Sarvaiya, S. Patnaik, and S. Bombaywala, "Image registration by template matching using normalized cross-correlation," in *Proc. Int. Conf. Adv. Comput., Control, Telecommun. Technol.*, Dec. 2009, pp. 819–822.
[38] J. P. Kern and M. S. Pattichis, "Robust multispectral image registration using mutual-information models," *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 5, pp. 1494–1505, May 2007.
[39] 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.
[40] X. Xu, X. Li, X. Liu, H. Shen, and Q. Shi, "Multimodal registration of remotely sensed images based on Jeffrey's divergence," *ISPRS J. Photogramm. Remote Sens.*, vol. 122, pp. 97–115, Dec. 2016.
[41] F. Ye, Y. Su, H. Xiao, X. Zhao, and W. Min, "Remote sensing image registration using convolutional neural network features," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 2, pp. 232–236, Feb. 2018.
[42] 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.
[43] 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, pp. 1–5, 2022, doi: 10.1109/LGRS.2021.3105567.
[44] H. Le, F. Liu, S. Zhang, and A. Agarwala, "Deep homography estimation for dynamic scenes," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 7652–7661.
[45] 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. 15950–15959.
[46] Z. Jiang, F.-F. Yin, Y. Ge, and L. Ren, "A multi-scale framework with unsupervised joint training of convolutional neural networks for pulmonary deformable image registration," *Phys. Med. Biol.*, vol. 65, no. 1, Jan. 2020, Art. no. 015011.
[47] X. Huang, J. Ding, and Q. Guo, "Unsupervised image registration for video SAR," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 1075–1083, 2021.
[48] 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.
[49] J. Hu, L. Shen, and G. Sun, "Squeeze-and-Excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.
[50] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[51] 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.
[52] A. Paszke *et al.*, "Pytorch: An imperative style, high-performance deep learning library," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 8026–8037.
[53] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2005, pp. 886–893.

Yuanxin Ye (Member, IEEE) received the B.S. degree in remote sensing science and technology from Southwest Jiaotong University, Chengdu, China, in 2008, and the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2013.

He is currently a Research Fellow with the Faculty of Geosciences and Environmental Engineering, Southwest Jiaotong University. His research interests include remote sensing image processing, image registration, change detection, and object detection.

Dr. Ye received the ISPRS Prizes for Best Papers by Young Authors of the 23rd International Society for Photogrammetry and Remote Sensing Congress, Prague, in 2016, and the Best Youth Oral Paper Award of ISPRS Geospatial Week 2017, Wuhan, in 2017.

Tengfeng Tang received the B.S. degree in geographic information science from the China University of Petroleum (East China), Qingdao, China, in 2020. He is currently pursuing the M.S. degree with the Faculty of Geosciences and Environmental Engineering, Southwest Jiaotong University, Chengdu, China.

His research interests include deep learning, remote sensing image processing, and image registration.

Bai Zhu received the B.S. degree from the Faculty of Geosciences and Environmental Engineering, Southwest Jiaotong University, Chengdu, China, in 2019, where he is currently pursuing the Ph.D. degree.

His research is mainly focused on multimodal image matching, image registration, and feature extraction.

Chao Yang received the B.S. degree from the Faculty of Geosciences and Environmental Engineering, Southwest Jiaotong University, Chengdu, China, in 2019, where he is currently pursuing the Ph.D. degree in surveying and mapping science and technology.

His research interests include image matching, deep learning, and image processing.

Bo Li received the B.S. and Ph.D. degrees from Northwestern Polytechnical University, Xi'an, China, in 2011 and 2018, respectively.

He is currently a Research Associate with Northwestern Polytechnical University. He has authored more than 20 articles in *Pattern Recognition* (PR), IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING (TGRS), and IEEE TRANSACTIONS ON MULTIMEDIA (TMM). His research interests include deep learning, deep reinforcement learning, and computer vision.

Siyuan Hao (Member, IEEE) received the Ph.D. degree from the College of Information and Communications Engineering, Harbin Engineering University, Harbin, China, in 2015.

She is currently an Associate Professor with the Qingdao University of Technology, Qingdao, China, where she teaches remote sensing and electrical communication. Her research interests focus on hyperspectral imagery processing and machine learning.