MU-Net: A MULTISCALE UNSUPERVISED NETWORK FOR REMOTE SENSING IMAGE REGISTRATION

Tengfeng Tang, Tingting Chen, Bai Zhu, Yuanxin Ye*

Faculty of Geosciences and Environmental Engineering, Southwest Jiaotong University, Chengdu 610031, China

Commission III, WG III/VI

KEY WORDS: Image Matching, Image Registration, Unsupervised Learning, Multi-Modal Images.

ABSTRACT:

Registration for multi-sensor or multi-modal 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 unsupervised network (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 performs a coarse-to-fine registration pipeline by stacking several deep neural network models on multiple scales, which prevents the backpropagation being falling into a local extremum and resists significant image distortions. In addition, a novel loss function paradigm is designed based on structural similarity, which makes MU-Net suitable for various types of multi-modal 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-SAR and Optical-Map datasets. Experimental results show that MU-Net achieves more robust and accurate registration performance between these image pairs with geometric and radiometric distortions. We share the datasets and the code implemented by Pytorch at https://github.com/yeyuanxin110/MU-Net.

1. INTRODUCTION

Remote sensing image registration (RSIR) aims to obtain geometric transformation parameters (TPs) between images by correspondence detection. RSIR is a preliminary task, which directly influences the performance of the following tasks, such as image fusion, change detection and deformation monitoring. From traditional to deep learning (DL) techniques, many inspiring methods for image registration have been developed in the remote sensing community.

Traditional methods can be generally classified into two categories: feature-based methods and area-based methods (Ma et al., 2021). Feature-based methods extract the salient and repeatable features from two images and establish their correspondences. The scale invariant feature transform (SIFT) (Lowe, 2004) is a representative, and many algorithms are proposed on the basis of SIFT, such as the speed up robust feature (SURF) (Bay, 2008), and the oriented FAST and rotated BRIEF (ORB) (Rublee, 2011). These SIFT-like methods are suitable for extracting repeatable features from single-modal images, but vulnerable to multi-modal images with radiometric changes such as optical-SAR image pairs. To improve the robustness to radiometric differences, some local feature descriptors were proposed, such as the radiation-variation insensitive feature transform (RIFT) (Li et al., 2020). Overall, the main challenge of feature-based methods is to extract highly repeatable and distinct features and match them correctly.

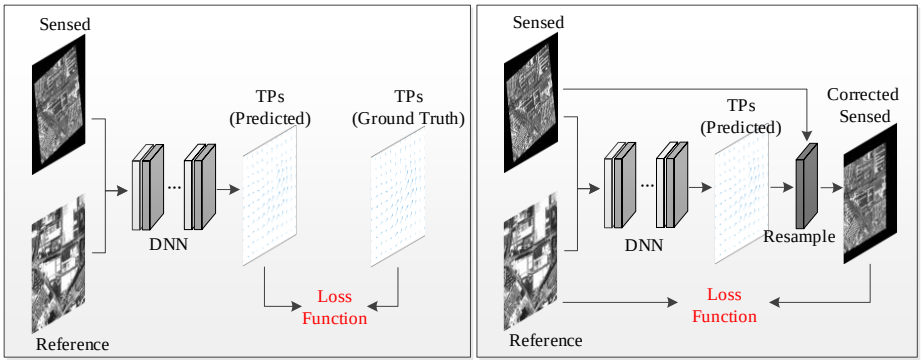
Area-based methods often adopt a template scheme and detect correspondences by evaluating the similarity of images. Some widely used similarity metrics are the sum of squared differences (SSD), the normalized cross correlation (NCC) (Sarvaiya et al., 2009) and the mutual information (MI) (Kern et al., 2007). The performance of the above metrics is easily affected by radiometric changes. The latest research of area-based methods is to integrate structural features into similarity metrics for coping with radiometric changes (Ye et al., 2017; Ye et al., 2019). Although the area-based methods based on structural features can effectively address radiometric changes, it is necessary to eliminate the obvious geometric distortions between images before image registration. This requires manual selection of control points or requires the images with geo-referenced information, making area-based methods limited. Overall, the main limitation of area-based methods is that they cannot effectively handle the images with large geometric distortions.

The two types of traditional methods respectively have the above-mentioned shortcomings. Besides, they commonly include a matching process integrating features or local descriptors extracted by handcraft rather than automated learning. Since no information feedback among feature extraction, description and matching, these approaches lack deep-level semantic information. When image 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 multi-modal images.

Recent years, a growing number of researches focus on DL. To a certain extent, DL methods can solve the shortcomings of traditional methods. Generally, DL methods can be classified into two categories: integrated learning methods and end-to-end learning methods (Jiang et al., 2020).

Integrated learning methods usually integrate a deep neural network (DNN) (Liu et al., 2017) into a traditional method, and extract feature descriptors from the auto-learned feature maps. Traditional operations like keypoint detection, feature description or template sliding are conducted on the auto-learned

* Corresponding author (Email: yeyuanxin@home.swjtu.edu.cn)



(a) (b)

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

feature maps instead of the original images or the handicraft feature maps. Some researches combined SIFT descriptor with a DNN (Ye et al., 2018), or put multi-orientated gradient features into a DNN (Zhou et al., 2022). However, these integrated learning methods cannot match images with large geometric distortions, and they still require designing a specific DNN for different data. Compared with traditional methods, the computational complexity increases many times, but the registration effect has not been improved significantly.

End-to-end learning methods aim to directly predict the TPs. According to whether the optimizer require the ground truth TPs, end-to-end learning methods can be divided into supervised end-to-end learning methods (hereafter called supervised methods), and unsupervised end-to-end learning methods (hereafter called unsupervised methods) (Jiang et al., 2020), and their common architecture are shown in Figure 1 (a) and (b), respectively.

Supervised methods minimize the discrepancy between the predicted TPs and the ground truth TPs during training process, as Figure 1 (a) shown. Related studies include the Deep Image Homography Estimation Network (DHN) (DeTone et al, 2016), the Multi-scale Deep Image Homography Estimation Network (MHN) (Le et al., 2020), and the Deep Lucas-Kanade Feature Map (DLKFM) (Zhao et al., 2021). However, the network of supervised methods requires training by a large number of images with the ground truth TPs. One big challenge is that true labels are costly and hard to acquire in RSIR. Such limitation makes supervised methods difficult to be widely applied in practice.

Unsupervised methods optimize the similarity between images during training process, and the ground truth TPs is not required, as Figure 1 (b) shown. Recently, unsupervised methods have been developed in medical image registration, because it solves the problem that network cannot be trained effectively with no ground truth TPs. Related masterpieces include VoxelMorph (Balakrishnan et al., 2019), and the deep learning image registration framework (DLIR) (Vos et al., 2019). However, it may be not appropriate to directly apply the related methods to RSIR for the following reasons. Firstly, current methods cannot effectively handle noise and non-linear radiometric differences, which make these methods vulnerable for multi-modal RSIR. Secondly, these methods require images roughly aligned before image registration, whereas in RSIR, it's the goal rather than a preprocessing step to eliminate the geometric distortions. When images have significant geometric and radiometric differences, these methods often suffer large registration errors.

Generally speaking, there is a lack of which can effectively and simultaneously handle the large geometric distortions and radiometric differences between images without the ground truth TPs, and our work has filled this gap.

We propose a multiscale unsupervised network (MU-Net) for RSIR, 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 between images, thereby circumventing the need for the ground truth TPs. Firstly, each DNN model is individually and successively trained to initialize the network weights. Secondly, all DNN models are stacked in a cascading way to form a combined registration pipeline, and the parameters of which are jointly trained to output the final TPs. Besides, the similarity evaluation of image pairs is performed on structural features rather than image intensity, which is suitable for multi-modal RSIR.

Our main contributions have three aspects:

(1) We propose a registration network with unsupervised learning, which is an end-to-end mapping scheme from the image pairs to their transformation parameters.
(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 resist a large range of image distortions.
(3) We design a novel loss function paradigm based on structural similarity, which makes the registration network suitable for various types of multi-modal images.

2. METHODOLOGY

In this section, the proposed MU-Net for RSIR is elaborated, which passes the images through several designed DNN architectures on multiple scales to regress the TPs, 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 two images, MU-Net is completely unsupervised. In this paper, we choose affine TPs as the form of the predicted mapping, and MU-Net can integrate other forms (e.g. homography and Deformable Vector Field (DVF) (Balakrishnan et al., 2019)) of TPs. The details are given in below.

2.1 Problem Formulation

Assuming there is a pair of images f & m to be aligned. One is a reference image f with correct geographic coordinates for each

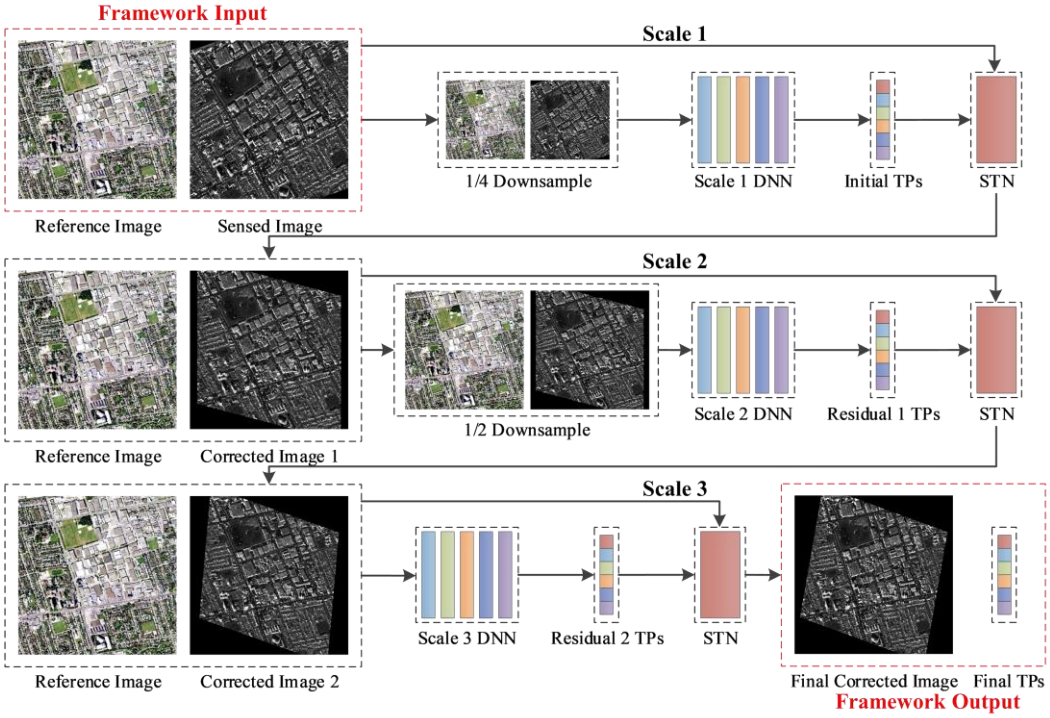


Figure 2. Multiscale workflow for MU-Net.

pixel, and the other is a sensed image m with geometric distortions. To correct m, the aim is to find a group of TPs μ.
In traditional image registration, μ is directly optimized by maximizing a certain similarity metric *Sim*:

$$\hat{\mu} = \text{argmax}_{\mu}[Sim(f, T_{\mu}(m))], \quad (1)$$

where T_{μ} means a coordinate or spatial mapping parameterized by μ, and $\hat{\mu}$ is the optimal value of μ.
In the unsupervised registration method, μ is regressed by the designed DNN F:

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

where θ refers to the weights and bias parameters of F.
Therefore, μ is optimized indirectly, since it's θ to be directly optimized by maximizing the *Sim*:

$$\hat{\theta} = \text{argmax}_{\theta}[\text{Sim}(f, T_{\mu}(m))], \quad (3)$$

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 is optimized during training process.

2.2 Multiscale Workflow

MU-Net performs a multiscale coarse-to-fine strategy. Specifically, three DNN models are stacked in a cascading way, and images of different down-sampling rate are input to MU-Net, as shown in Figure 2.

Firstly, the DNN model in scale 1 performs an initial and global alignment between the input images f & m. Specifically, f & m are down-sampled by a scale factor of 1/4, then input into the first DNN model to evaluate the initial TPs μ_1. Subsequently, μ_1 is applied to a Spatial Transformer Network (STN) (Max et al., 2016), and to correct the original sensed image m to produce the first corrected sensed image $T_{\mu 1}(m)$.

Secondly, the DNN model in scale 2 performs a residual alignment between f & $T_{\mu 1}(m)$. Specifically, f & $T_{\mu 1}(m)$ are down-sampled by a scale factor of 1/2, then input 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 a STN and to correct the original sensed image m to produce the second corrected sensed image $T_{\mu 2}(m)$.

Thirdly, the DNN model in scale 3 also performs a more detailed alignment between f & $T_{\mu 2}(m)$. Specifically, f & $T_{\mu 2}(m)$ are directly input to the third DNN model to evaluate the residual TPs $\Delta\mu_2$, which is integrated to μ_2 to yield the final TPs μ_3. And μ_3 is applied to a STN and to correct the original sensed image m to produce the final corrected sensed image $T_{\mu 3}(m)$, thereby achieving the image registration.

2.3 DNN Architecture on Each Scale

In this section, we describe the DNN architecture on each scale. In order to extract the deep semantic information and find the end-to-end TPs mapping, we utilize the channel attention mechanism (Hu et al., 2018) and the deep residual network (He et al., 2016) to form the DNN architectures. The former can adaptively adjust the weight of each channel. And the latter ensures that the deep semantic information will not decrease as the network deepening.

We define a Deep Residual (DR) ConvBlock as an ordinary convolution block with a residual network added, and a (Squeeze Excitation and Deep Residual) SE-DR ConvBlock means a DR ConvBlock with the channel attention mechanism integrated. Figure 3 depicts the DNN architecture on the third scale. Input image pairs should have the same size, 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

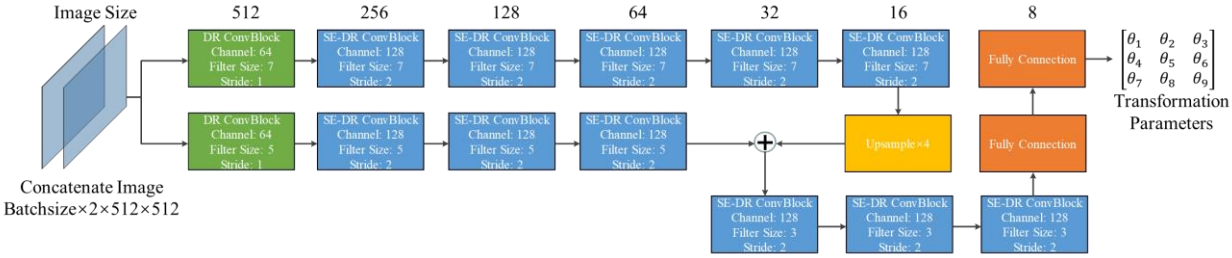


Figure 3. Architecture of DNN architecture in scale 3.

ConvBlocks and a series of 5×5 ones, respectively. The two routes are concatenated by up-sampling and stride connection, followed by several 3×3 ConvBlocks. During the forward propagation, the image size is reduced while the channel deepens, which is conducive to extract the deep semantics information. After through the last ConvBlock, the deep semantic information is directly mapped to the TPs through two fully connection layers.

The DNN architectures on the first and second scales are similar to the one on the third scale. Where as the difference is that, the input image has undergone a down-sampling, so the initial size of the image has become 128×128 pixels or 256×256 pixels instead of 512×512 pixels. Therefore, we reduce two SE-DR ConvBlocks for each route, while maintaining the maximum number of channels at 32, which forms the DNN architecture on the first scale. Similarly, we reduce a SE-DR ConvBlock for each route and maintain the maximum number of channels at 64, which forms the DNN architecture on the second scale.

2.4 Unsupervised Training

In MU-Net, three DNN models are stacked in a cascading way to form a coarse-to-fine registration pipeline. Therefore, the training procedure includes two parts, initialization and joint training.

In the first stage, to initialize the network weights, each DNN model was individually and successively trained, to minimize the corresponding loss based on image structural similarity: $Loss_{sim}(f, m, \mu_1)$, $Loss_{sim}(f, m, \mu_2)$ and $Loss_{sim}(f, m, \mu_3)$. The first model was trained for a rough alignment. With weights fixed for the first model, the second model was successively trained to fine-tune the alignment. Finally, the third model e was trained to further correct the alignment, while freezing the weights of the first and second models.

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

$$\begin{aligned} Loss = \lambda_1 Loss_{sim}(f, m, \mu_1) + \\ \lambda_2 Loss_{sim}(f, m, \mu_2) + \lambda_3 Loss_{sim}(f, m, \mu_3) \end{aligned}, \quad (4)$$

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

The reference image is supported to achieve the best similarity with the sensed image corrected by the TPs and its spatial transformation. Similarly, the sensed image is supported to achieve the best similarity with the reference image wrapped by the inverse spatial transformation. To improve the reliability of the TPs μ, we invert the matrix of the coordinate mapping T_μ:

$$T_\mu^{-1} = (T_\mu^T T_\mu)^{-1} T_\mu^T, \quad (5)$$

where T_μ^{-1} denotes the inversed matrix of the coordinate mapping. Therefore, the similarity loss function is defined as:

$$Loss_{sim}(f, m, \mu) = e^{-[Sim(f, T_\mu(m)) + Sim(T_\mu^{-1}(f), m)]/2}, \quad (6)$$

where e is the natural constant, and *Sim* is a certain similarity metric. A higher *Sim* denotes a better similarity and a lower $Loss_{sim}$.

For multi-modal RSIR, such as optical-SAR images, their pixel intensity cannot be directly used for similarity evaluation due to radiometric differences. Considering that structure features are preserved between multi-modal images, we use the structural descriptor instead of intensity to calculate the value of similarity metric. To convergence loss function, we mainly adopted a fast and robust structural descriptor named the channel features of orientated gradients (CFOG) (Ye et al., 2019). As Fig.6 shown, CFOG first extracts multi-oriented gradients and then construct the oriented histogram. Based on oriented histogram, the convolution operation is performed by a 3-D Gaussian-like

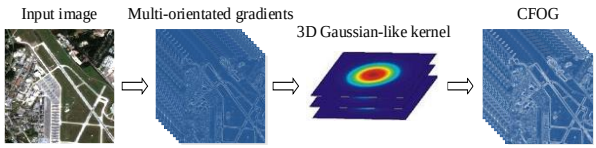


Figure 4. Construction process of CFOG

kernel which collects the orientated gradients of neighbouring pixels. Thus, a 3-D structural feature map is generated.

We adopt the similarity metric NCC for *Sim*(*A*, *B*) on the structural feature maps *A* and *B*. NCC determines the correspondences between two structural feature maps by searching the location of maximum value, which can be computed as:

$$NCC(A, B) = \frac{\sum_{p \in N}(A(p) - \tilde{A})(B(p) - \tilde{B})}{\sqrt{\sum_{p \in N}(A(p) - \tilde{A})}\sqrt{\sum_{p \in N}(B(p) - \tilde{B})}}, \quad (7)$$

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 denotes a higher similarity.

3. EXPERIEMENT

3.1 Datasets and Comparison Methods

3.1.1 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. As an example shown in Figure 5, I_1 and I_2 are two precisely aligned images with a size of

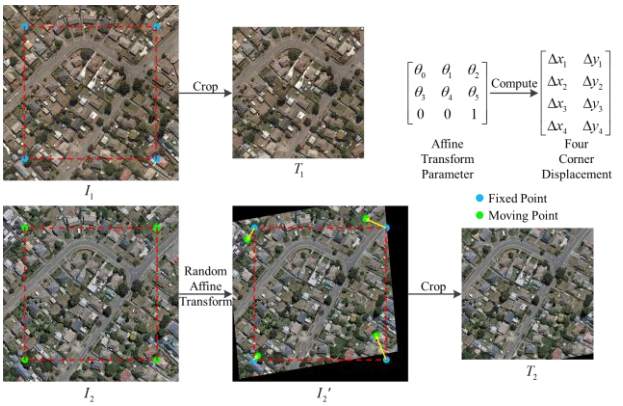


Figure 5. An example for generation of an image pairs.

larger than 512×512 pixels. I_2 is randomly performed affine transformation to generate wrapped image I_2'. Then, the same positions of I_1 and I_2' are cropped, which obtain reference image T_1 and sensed image T_2 as the image pair with a 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:

$$\begin{bmatrix} \Delta x_i \\ \Delta y_i \\ 1 \end{bmatrix} = \left(\begin{bmatrix} \theta_0 & \theta_1 & \theta_2 \\ \theta_3 & \theta_4 & \theta_5 \\ 0 & 0 & 1 \end{bmatrix} - E \right) \begin{bmatrix} x_i \\ y_i \\ 1 \end{bmatrix}, \tag{8}$$

where $[x_i, y_i]$ represents the position of the *i*-th 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 subsequent experiments. The specific experimental datasets will be introduced in detail in the following subsection.

3.1.2 Datasets

MU-Net has been extensively evaluated on the Optical-Optical, Optical-Infrared, Optical-SAR and Optical-Map datasets. These datasets are composed of image pairs generated according to Section 3.1.1, and they are adopted for two purposes. In particular, since SAR images contain obvious speckle noise and raster maps contain text labels, the registration tasks on the Optical-SAR and Optical-Map datasets are more challenging than other cases.

Optical-Optical. WHU Building Dataset is a set of multi-temporal aerial images with a resolution of 0.075m in Christchurch, New Zealand, in 2012 and 2016. We crop these areas into about 500 image pairs with a size of 600×600 pixels. For each pair of images, we perform 10 random affine transformations and centre crops on them. The Optical-Optical dataset includes about 5000 image pairs with a size of 512×512 pixels. Approximately 4,500 pairs are used for training, and 500 for testing.

Optical-Infrared. The Optical-Infrared dataset includes the Chengdu Plain with a resolution of 30m. 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. Similar to the above operation, approximately 4,500 pairs are used for training, and 500 for testing.

Optical-SAR. The coverage of Optical-SAR dataset includes difference scenes such as cities, farmland, rivers and forests. The image pair are acquired by Sentinel-1 and Sentinel-2 in May 2021 with a resolution of 10m. Around 5,000 pairs are used for training, and 800 for testing.

Optical-Map. We obtained the optical images and the corresponding Google maps with a resolution of 1m from the Google map service. The image area is located is in Tokyo, and the ground objects are mainly dense buildings and streets. The maps of these scenes and the corresponding optical images look structural similar. Around 5,000 pairs are used for training, and 800 for testing.

3.1.3 Comparison Methods

As the registration difficulty of the four datasets is gradually increasing, we could observe the trend of the performance of MU-Net compared with state-of-the-arts methods [i.e. SIFT, RIFT, CFOG, DLKFM and DLIR], and evaluate their flexibility on different types of datasets. Among them, SIFT and RIFT are traditional handicraft methods with feature matching, CFOG is a traditional handicraft method with template matching, DLKFM is a DL method with supervised end-to-end mapping, and DLIR and MU-Net are DL methods with unsupervised end-to-end mapping.

3.2 Evaluation Criteria and Implementation Details

3.2.1 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, equation (8) is adopted to transform the affine TPs into the four corner displacements, 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 (DeTone et al, 2016; Le et al., 2020; Zhao et al., 2021), 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 $[\Delta x_i^{gt}, \Delta y_i^{gt}]$ and the predicted ones $[\Delta x_i^{pre}, \Delta y_i^{pre}]$. Note that these ground truth TPs are merely used for accuracy evaluation, and are not utilized during the training process.

3.2.2 Implementation Details

Our experiments adopted the following settings. The NCC on the CFOG feature maps is adopted as the similarity metric of MU-Net. Random affine transformations on sensed images consist of 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 value is limited in a range of [-0.1, 0.1] with a precision of 0.002 (about 1 pixel in an image with size of 512×512 pixels), the rotation angle is limited in a range of [-π, π] with a precision of 1 degree and the shear angle is limited in a range of [-π/6, π/6] with a precision of 1 degree. 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]. In MU-Net, all DNN models predict 6 affine TPs.

For compared methods, we use their settings recommended by the authors. In our experiments, the traditional matching methods (i.e. SIFT, RIFT and CFOG) adopt Random Sample Consensus (RANSAC) (Fischler, Bolles, 1981) to fit the affine TPs.

3.3 Accuracy Analysis

In this section, we use the test datasets 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. For each method tested on different datasets, Table 1 lists the percentage of the test images with an ACE smaller than 3 pixels. Figure 6 and 7 show two examples. Generally speaking, on the four multi-temporal or multi-modal datasets, MU-Net achieves the best registration accuracy.

Dataset	Optical-Optical	Optical-Infrared	Optical-SAR	Optical-Map
MU-Net	96%	96%	92%	89%
SIFT	51%	39%	0%	0%
RIFT	63%	54%	53%	17%
CFOG	15%	14%	14%	12%
DLKFM	92%	63%	59%	56%
DLIR	82%	1%	0%	0%

Table 1. The percentage of test images with the ACE smaller than 3 pixels for comparing mentioned methods.

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. Figure 8 show the ACE of each method for translation, rotation and scale deformation, respectively. Similarly, these experiments are performed on different multi-modal datasets. We can observe and evaluate the registration performance from simple to difficult data. Accordingly, further discussions about each method are given below.

SIFT. This is one of the classic feature matching methods. From Table 1, SIFT achieve 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 Optical-Map datasets. From Figure 8, SIFT is robust to translation, scale, and rotation differences between single-modal images (e.g. Optical-Optical), but it is vulnerable to multi-modal image matching (e.g. Optical-SAR).

RIFT. This is a feature matching method for multi-modal remote sensing images. From Table 1, there is no significant change in the registration performance on different datasets. As can be seen from Figure 8, RIFT has a good invariance for translation and rotation, but it cannot handle images with scale changes.

CFOG. This is a template matching method with fast computational efficiency and robust matching performance. From Figure 8 (a), in the case of only translation distortions, the registration performance of CFOG achieves the best, and it is robust to radiometric differences for multi-modal images with inconsistent translations. Whereas Figure 8 (b) and (c) show that CFOG is sensitive to rotation and scale differences.

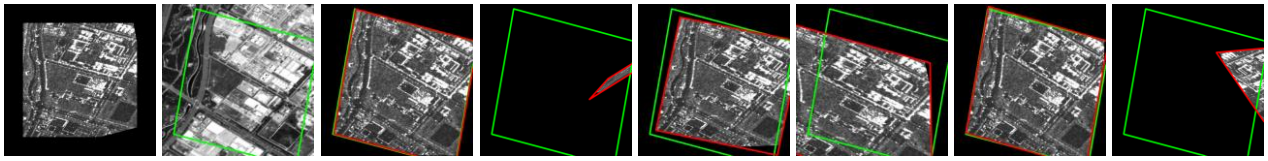
DLKFM. This method adopts supervised end-to-end learning, which trains its network with the L2 distance between the predicted TPs and the ground truth TPs as the loss function. From Table 1, DLKFM aligns 92% of test image pairs on Optical-Optical images within a 3-pixel ACE. Nevertheless, its performance on the three other multi-modal datasets has dropped significantly. On these datasets, the percentage of ACE within 3 pixels are drastically reduced, indicating that the radiometric resistance learned by DLKFM has a certain limitation. From Figure 8, DLKFM is robust to translation and rotation differences, but easily influenced by radiometric changes.

DLIR. This is a pioneering method of unsupervised learning in medical imaging. Since this method is to firstly predict affine TPs and finally predict a DVF for a dense matching, we only adopted the first network introduced in the author's literature for predicting affine TPs. From Table 1, DLIR aligns 82% of test image pairs on Optical-Optical datasets within a 3-pixel ACE, but it achieves poor performance for the other three multi-modal datasets. From Figure 8, 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 1, MU-Net obtains the best accuracy, and the accuracy is almost not affected by different image modals. For example, on the Optical-Optical dataset that are the easiest to be aligned, MU-Net aligns 96% of test image pairs within a 3-pixel ACE, and aligns 89% ones even on the Optical-Map dataset that are the most difficult for image registration. From Figure 8 (a), on the test images of all modalities, the accuracy of MU-Net

(a) (b) (c) (d) (e) (f) (g) (h)

Figure 6. Registration Result of an image pair in Optical-Optical dataset. (a) Randomly affine transformed and cropped sensed image. (b) Reference image. (c) MU-Net. (d) SIFT. (e) RIFT. (f) CFOG. (g) DLKFM. (h) DLIR. The green line represents the ground truth registration result, and the red line represents the experimental registration result for each method.



(a) (b) (c) (d) (e) (f) (g) (h)

Figure 7. Registration Result of an image pair in Optical-SAR dataset. (a) Randomly affine transformed and cropped sensed image. (b) Reference image. (c) MU-Net. (d) SIFT. (e) RIFT. (f) CFOG. (g) DLKFM. (h) DLIR. The green line represents the ground truth registration result, and the red line represents the experimental registration result for each method.

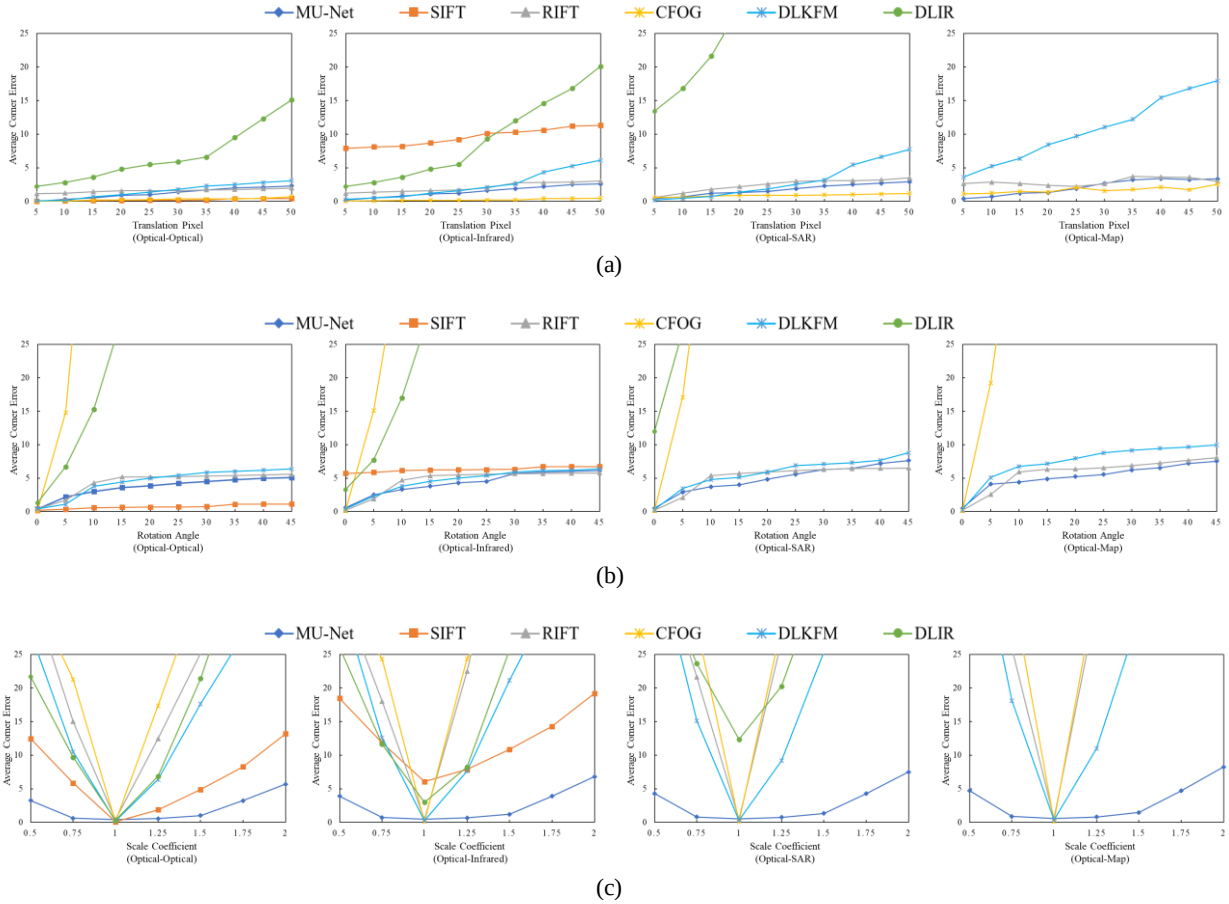


Figure 8. Comparison of ACE on different datasets for (a) translation distortion, (b) rotation distortion, and (c) scale distortion.

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 Figure 8 (b), on the Optical-Optical and Optical-Infrared datasets, MU-Net is comparable to the supervised DLKFM, and outperforms the RIFT. As the registration difficulty increases, MU-Net is least affected by rotation changes on the Optical-SAR and Optical-Map datasets, and presents the best registration performance. It can be clearly observed from Figure 8 (c) 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 multi-modal image registration tasks with scale changes. On the four 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 5 pixels.

The above experimental results prove that MU-Net can be applied to various types of multi-modal 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.

3.4 Analysis of Noise Sensitivity

This section examines the noise sensitivity of MU-Net, comparing it with SIFT, DLKFM and DLIR. The percentage of the test images within a 3-pixel ACE is used to analyse the noise sensitivity. For each image pairs on the Optical-Optical test dataset, 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. The reason to choose the above methods on the Optical-Optical dataset for comparison is, the experimental results in section 3.3 have shown that the above methods can handle certain affine distortions on the Optical-Optical dataset. On the other hand, there is no significant radiometric differences on the Optical-Optical dataset, which makes the experiment can objectively evaluate the influence of noise for these methods.

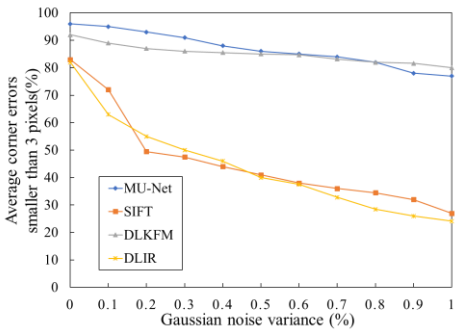


Figure 9. Percentage of images with ACE smaller than 3 pixels versus various Gaussian noise.

Figure 9 shows the percentage of ACE <= 3 pixels versus various Gaussian noise in image registration. Under Gaussian noise, MU-Net and DLKFM performs better than other methods, which indicates that MU-Net can achieve a comparable performance of the supervised method. Although SIFT can handle affine distortions between single-modal images, it is less able to resist Gaussian noise. As an unsupervised method, DLIR presents the highest noise sensitivity, whereas MU-Net overcomes this shortcoming

.

4. CONCLUSION

In this paper, we propose a multiscale unsupervised network (MU-Net) for remote sensing image registration. Without the ground truth labels, MU-Net directly learns the end-to-end mapping from the image pairs to their transformation parameters. MU-Net stacks several DNN models on multiple 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 multi-modal images. Experiments are performed on four datasets, including the Optical-Optical, Optical-Infrared, Optical-SAR and Optical-Map. Experimental results show that MU-Net is more robust to geometric and radiometric distortions between multi-modal images and achieves higher registration accuracy, compared with the current state-of-the-art traditional methods (such as SIFT, CFOG and RIFT), supervised and unsupervised deep learning methods (such as DLIR and DLKFM).

MU-Net is flexible for remote sensing image registration. This is because it can regress the parameters of various transformation models (e.g. homography and DVF) beside the used affine model. Moreover, other structure feature descriptors such as the histogram of oriented phase congruency (HOPC) (Ye et al., 2017) and the histogram of oriented gradient (HOG) (Dalal et al., 2005) can also be integrated as similarity metrics for calculating the loss function in MU-Net.

REFERENCES

Balakrishnan, G. et al., 2019. Voxelmorph: A learning framework for deformable medical image registration. *IEEE Transactions on Medical Imaging*, 38, 8, 1788-1800.

Bay, H. et al., 2008. Speeded-up robust features (SURF). *Computer Vision and Image Understanding*, 110(3), 346-359.

Dalal, N. et al., 2005, Histograms of oriented gradients for human detection, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 886–893.

DeTone, D. et al., 2016. Deep image homography estimation. *Computer Vision and Pattern Recognition*, arXiv: 1606.03798.

Fischler, M. A., Bolles, R. C., 1981. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6), 381-395.

He, K. et al., 2016. Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770-778.

Hu, J. et al., 2018. Squeeze-and-Excitation Networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 7132-7141.

Jiang, Z. et al., 2020. A multi-scale framework with unsupervised joint training of convolutional neural networks for pulmonary deformable image registration. *Physics in Medicine & Biology*, 65(1), 015011.

Kern, J. P. et al., 2007. Robust multispectral image registration using mutual-information models, *IEEE Transactions on Geoscience and Remote Sensing*, 45(5), 1494–1505.

Le, H. et al., 2020. Deep homography estimation for dynamic scenes. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 7652-7661.

Li, J. et al., 2020. RIFT: Multi-modal image matching based on radiation-variation insensitive feature transform. *IEEE Transactions on Image Processing*, 29, 3296–3310.

Liu, W. et al., 2017. A survey of deep neural network architectures and their applications. *Neurocomputing*, 234, 11-26.

Lowe, D. G., 2004. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2), 91-110.

Ma, J. et al, 2021. Image matching from handcrafted to deep features: A survey. *International Journal of Computer Vision*, 129, 23-79.

Max, J. el al, 2016. Spatial Transformer Networks. *Computer Vision and Pattern Recognition*, arXiv:1506.02025.

Rublee, E. et al., 2011. ORB: An efficient alternative to SIFT or SURF. *International Conference on Computer Vision*, 2564-2571.

Sarvaiya, J. N. et al., 2009. Image registration by template matching using normalized cross-correlation. *International Conference on Advances in Computing, Control, and Telecommunication Technologies*, 819-822.

Vos, B. D. et al., 2019. A deep learning framework for unsupervised affine and deformable image registration. *Medical Image Analysis*, 52, 128-143.

Ye, F. et al., 2018. Remote sensing image registration using convolutional neural network features. *IEEE Geoscience and Remote Sensing Letters*, 15(2), 232-236.

Ye, Y. et al., 2017. Robust registration of multi-modal remote sensing images based on structural similarity, *IEEE Transactions on Geoscience and Remote Sensing*, 55(5), 2941-2958.

Ye, Y. et al., 2019. Fast and robust matching for multi-modal remote sensing image registration, *IEEE Transactions on Geoscience and Remote Sensing*, 57(11), 9059-9070.

Zhao, Y. et al., 2021. Deep Lucas-Kanade homography for multimodal image alignment. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 15950-15959.

Zhou, L. et al., 2021. Robust matching for SAR and optical images using multiscale convolutional gradient features. *IEEE Geoscience and Remote Sensing Letters*, 19, 1-5.