MURF: Mutually Reinforcing Multi-Modal Image Registration and Fusion

Han Xu, Jiteng Yuan, and Jiayi Ma, *Senior Member, IEEE*

***Abstract*—Existing image fusion methods are typically limited to aligned source images and have to "tolerate" parallaxes when images are unaligned. Simultaneously, the large variances between different modalities pose a significant challenge for multi-modal image registration. This study proposes a novel method called MURF, where for the first time, image registration and fusion are mutually reinforced rather than being treated as separate issues. MURF leverages three modules: shared information extraction module (SIEM), multi-scale coarse registration module (MCRM), and fine registration and fusion module (F2M). The registration is carried out in a coarse-to-fine manner. During coarse registration, SIEM first transforms multi-modal images into mono-modal shared information to eliminate the modal variances. Then, MCRM progressively corrects the global rigid parallaxes. Subsequently, fine registration to repair local non-rigid offsets and image fusion are uniformly implemented in F2M. The fused image provides feedback to improve registration accuracy, and the improved registration result further improves the fusion result. For image fusion, rather than solely preserving the original source information in existing methods, we attempt to incorporate texture enhancement into image fusion. We test on four types of multi-modal data (RGB-IR, RGB-NIR, PET-MRI, and CT-MRI). Extensive registration and fusion results validate the superiority and universality of MURF.**

***Index Terms*—Multi-modal images, image registration, image fusion, contrastive learning.**

I. Introduction

DUE to the limitations of hardware devices, images from one type of sensor can merely characterize partial information. For instance, the reflected light information captured by visible sensors can describe scene textures while it is susceptible to light and shading. Complementarily, the thermal radiation information captured by infrared sensors is insensitive to light and can reflect the essential attributes of scenes and objects. Multi-modal image fusion aims to synthesize a single image by integrating complementary source information from different types of sensors. As shown in Fig. 1, the single fused image exhibits better scene representation and visual perception, which can benefit various subsequent tasks, such as semantic segmentation [1], object detection, and tracking [2], scene understanding [3], *etc*. Therefore, image fusion has a wide variety of applications, from security to industrial and civilian fields [4], [5].

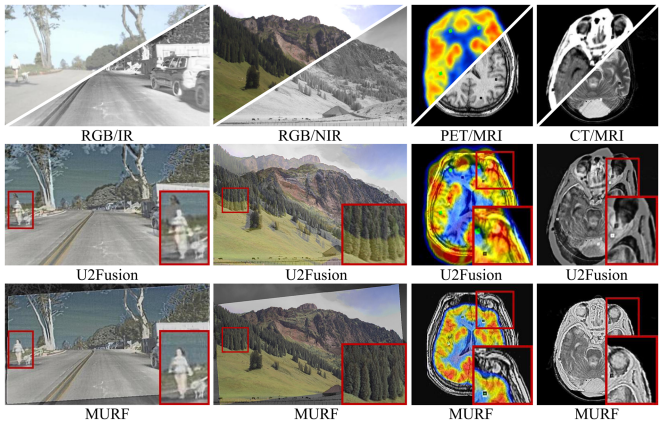


Fig. 1. Schematic illustration of different unaligned multi-modal image fusion tasks (first row: unaligned source images; second row: fusion results of a state-of-the-art fusion method (U2Fusion [6]); last row: results of MURF).

However, *existing fusion methods require source images to be accurately aligned and do not account for parallaxes*. When source images are unaligned, the parallaxes will lead to parallax fusion artifacts, as intuitively illustrated in the second row of Fig. 1. For unaligned source images, these fusion methods require other multi-modal image registration methods as pre-processing to eliminate parallaxes. In this case, registration and fusion are separate issues. As image fusion is merely a downstream task, the fused image cannot provide feedbacks to improve registration accuracy. Thus, existing fusion methods have to "tolerate" rather than "fight" the pre-registration misalignments, as illustrated in Fig. 2, which was also presented in our preliminary version [7]. Nevertheless, considering the characteristics of fused images, *it is possible for image fusion to inversely eliminate misalignments*. First, the fused images integrate the information from both modalities. The alleviated modal variances reduce the registration difficulty. Second, the fusion process discard some superfluous information, reducing its negative impact on registration. Third, misalignments in fused images lead to repeated salient structures, while accurate registration encourages gradient sparseness. Thus, gradient sparsity can act as a criterion to improve registration accuracy in a

Manuscript received 6 October 2022; revised 27 April 2023; accepted 5 June 2023. Date of publication 7 June 2023; date of current version 5 September 2023. This work was supported in part by the National Natural Science Foundation of China under Grant 62276192, and in part by the Key Research and Development Program of Hubei Province under Grant 2020BAB113. Recommended for acceptance by J. Zhou. *(Corresponding author: Jiayi Ma.)*

The authors are with the Electronic Information School, Wuhan University, Wuhan 430072, China (e-mail: xu_han@whu.edu.cn; yuanjiteng@whu.edu.cn; jyma2010@gmail.com).

Our code is publicly available at https://github.com/hanna-xu/MURF.

Digital Object Identifier 10.1109/TPAMI.2023.3283682

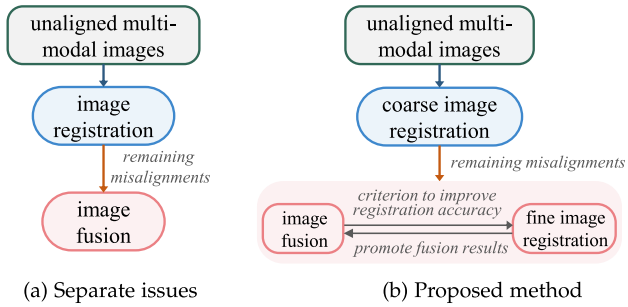


Fig. 2. Separate registration and fusion in existing methods and the proposed mutually reinforcing framework.

feedback fashion. When image fusion helps eliminate misalignments, the more precisely aligned data further promotes fusion results.

Specific to individual tasks, either multi-modal image registration or fusion has its own bottlenecks. *For multi-modal image registration*, there are three remaining challenges. First, *it is difficult to design a registration method applicable to multi-modal data that can break through the barriers of modal variances*. Existing metrics are insensitive to modal variances. Some metrics assume that the intensity distributions of multi-modal images have linear correlation, which is not always the case. Some methods use image translation to generate pseudo-mono-modal images while this method is contrary to the fact that multi-modal data is not one-to-one corresponding. Second, to eliminate modal variances through a transformation model, *some factors that may limit the implementation of transformation model should be considered*. For instance, we should consider the non-sparsity of features to enable network convergence, the computational complexity of loss function to facilitate the back propagation, and the possibility of existence of optimal solution. These factors make design of transformation model challenging. Third, *it is of practical importance to improve the registration universality.* Some methods only work for specific multi-modal data; some can only handle rigid deformations; and some methods effective for non-rigid deformations have difficulty in maintaining the rigidity of objects. Thus, it is necessary to design a method broadly applicable to a wide range of multi-modal data and both rigid and non-rigid deformations. *For image fusion*, a general purpose is to generate a single fused image to present the most amount of information, partly represented by gradients. Thus, *the preservation of scene content, especially textures, is an issue that most fusion methods strive to address.* In terms of practical applications, the fused image should contain more scene content and contribute positively to subsequent tasks. From this point of view, *it is reasonable and necessary to enhance low-quality texture details in source images into the fused image, rather than merely preserving original textures*. Unfortunately, this issue has not been noticed and addressed in existing fusion methods.

The proposed MURF addresses the limitations of existing multi-modal image registration methods and fusion methods by jointly realizing them in a mutually reinforcing framework. The proposed MURF consists of three main modules for shared information extraction, global rigid and local non-rigid deformation correction, and image fusion. The multi-modal image registration is handled in a coarse-to-fine approach. The coarse registration is based on the extracted mono-modal information and modeled as an affine transformation and realized through a multi-scale registration network. Fine registration and image fusion are realized in a single module, which relies on the characteristics of fused images to further improve registration accuracy and incorporates texture enhancement. The characteristics and contributions of MURF are summarized as:

1) To break through the bottleneck of requiring aligned source images in existing fusion methods, we for the first time interact multi-modal image registration and fusion in a mutually reinforcing framework through neural networks. Subsequently, the proposed method is applicable to unaligned source images, resulting in improved registration accuracy and fusion performance;
2) For multi-modal image registration, we apply a coarse-to-fine strategy where both global rigid transformations and local non-rigid transformations are considered. In the coarse phase, we transfer multi-modal image registration to the mono-modal shared information registration through contrastive learning. It enables application of metrics insensitive to modal variances. In the fine phase, the feedback of fused image and the explored inverse deformation both help correct misalignments;
3) For image fusion, we aim to not only preserve the scene content in original source images, but also to enhance their textures in the fused image for a more detailed scene expression. For these purposes, we design a fusion loss based on gradient evaluation, preservation, and enhancement, and introduce a gradient channel attention mechanism;
4) The proposed registration and fusion networks are applied to a variety of multi-modal data. We test the proposed MURF on four publicly available datasets, including RGB-IR, RGB-NIR, PET-MRI, and CT-MRI image pairs. Qualitative and quantitative results validate the universality and superiority of MURF in terms of registration accuracy and fusion performance.

A preliminary version of this paper is RFNet [7]. The most significant new contribution is the broadening of application scenarios. In the preliminary version, limited by image translation, RFNet can only be applied to RGB-NIR image pairs of street scenes. In this version, by revising the method for eliminating modal variances between multi-modal images, MURF is applicable to more multi-modal combinations, including RGB-IR, RGB-NIR, PET-MRI, and CT-MRI image pairs. The specific technical improvements over the preliminary version are in four aspects:

1) For the way to eliminate modal variances (i.e., transform multi-modal registration into mono-modal registration), RFNet employs image translation. Its application scenarios are limited due to the lack of one-to-one correspondence in multi-modal data. MURF applies contrastive learning to extract shared information and transfers mono-modal from the image domain to a common feature domain, lifting the restriction of image translation;

2) For the coarse multi-modal image registration, we revise the single-scale registration to a multi-scale progressive registration strategy. It speeds up the convergence and increases the registration accuracy;
3) For fine registration, RFNet only relies on the feedback and properties of fused images to correct local offsets. In this work, on the basis of feedback, we also explore the inverse deformation field for supervision. It further improves the fine-registration accuracy;
4) For image fusion, RFNet designs network architecture and loss function for texture retention. In MURF, we aim not only to preserve the original textures but also to enhance the textures with poor visibility in source images and display the enhanced textures in the fused image to provide a more detailed scene expression.

II. Related Work

A. Multi-Modal Image Registration

Multi-modal image registration is complicated due to appearance variability caused by different modalities. The problem arises in multi-sensor and medical images [8]. Generally, image registration methods can be divided into traditional and deep learning-based methods.

Traditional methods can be classified into three categories [9]. *i)* Transformation-based methods. They manually analyse common characteristics of multi-modal images and extract descriptors to exhibit greater consistency [10], [11], [12]. Considering accuracy and optimization, the transformation models should be meticulously designed; *ii)* Measure-based methods. The correlation-based metrics [13], [14], [15] assume that the intensity distributions of multi-modal images have a linear correlation. They are only applicable to specific types of data, which limits their scope of application. For the information theory-based metrics [16], [17], [18], it is difficult to determine the global maximum in the entire space; *iii)* Optimization methods. They consider applicable structures and objective functions. Despite their good performance, traditional methods need to consider many factors when dealing with complicated multi-modal data.

To avoid the concerns of traditional methods, deep learning has been partly or wholly applied in transformation and optimization. On the one hand, some methods use networks to design transformation models. FIRE [19] uses a network to learn a modality-independent latent representation for cycle-consistent cross-modality synthesis. Similarly, Nemar [20] learns a cross-modal translation. Then, it optimizes a uni-modal metric comparing the translated transformed image and the other source image for registration. However, multi-modal data is not one-to-one corresponding. For instance, the same thermal radiation property in an infrared image can correspond to diverse textures in a visible image. *Thus, the reliance on cross-modal translation to eliminate modal variances is inaccurate*. On the other hand, deep learning-based methods also require metrics for optimization. Although deep learning can alleviate the optimization difficulties of traditional methods, the solution in these methods is still thorny. First, some information theory-based metrics are computationally intractable for gradient descent. Second, metrics are insensitive to modal variances while existing cross-modal translation is inaccurate. Third, some sparse common features are not suitable for gradient descent. It is difficult to achieve a balance between alleviating modal variances and ensuing feature richness. Therefore, *it remains challenging to develop appropriate registration metrics practical for multi-modal data and gradient descent.*

In this work, we apply a coarse-to-fine strategy for multi-modal registration. In the coarse phase to correct global parallaxes, we translate two source domains to a common space, thereby eliminating modal variances and transforming multi-modal registration into mono-modal registration. With features with higher consistency, the metrics insensitive to modal variances are applicable for optimization through gradient descent. In the fine phase to correct local offsets, we utilize the feedback of fused images and explore inverse deformation fields to further improve accuracy.

B. Multi-Modal Image Fusion

Existing fusion methods are tailored to aligned multi-modal image pairs. Traditional fusion methods include six categories: methods based on multi-scale transformation [21], [22], sparse representation [23], [24], subspace [24], [25], saliency [26], [27], hybrid methods [28], [29], and others [30], [31]. These methods are devoted to designing decomposition ways and fusion rules for the decomposed components in a manual way. However, the detailed and diverse manual designs make these methods complex, time-consuming, and laborious. Moreover, the manually designed approach usually forces the same decomposition ways for multi-modal images while ignoring the modal variances. Thus, traditional methods usually show limited fusion performances.

To solve drawbacks of traditional methods, some learning-based methods have been proposed. According to the network architecture, these methods can be divided into methods based on auto-encoder [32], [33], [34], convolutional neural networks (CNN) [35], [36], [37], and generative adversarial networks (GANs) [38], [39], [40], [41]. Some methods only preserve textures according to the modality (i.e., preserve textures of a specific modality while ignoring textures in the other source image). It results in the distortion of scene information. The GAN-based methods usually suffer from fake and blurred details. Moreover, these methods are limited to retaining the information in source images. Some methods attempt to provide more textures in the fusion results. [42] tried to fuse details of multi-spectral and panchromatic images into fused images. However, the upper limit of preserving these textures is still the original source information which has not been further improved. [43] can sharpen boundaries between different image objects while the improved information is of particular types, e.g., boundaries.

In this work, we focus on texture preservation and enhance the original textures with poor visibility. In terms of the network architecture, we introduce the gradient channel attention block to adaptively adjust the channel-wise contributions of features.

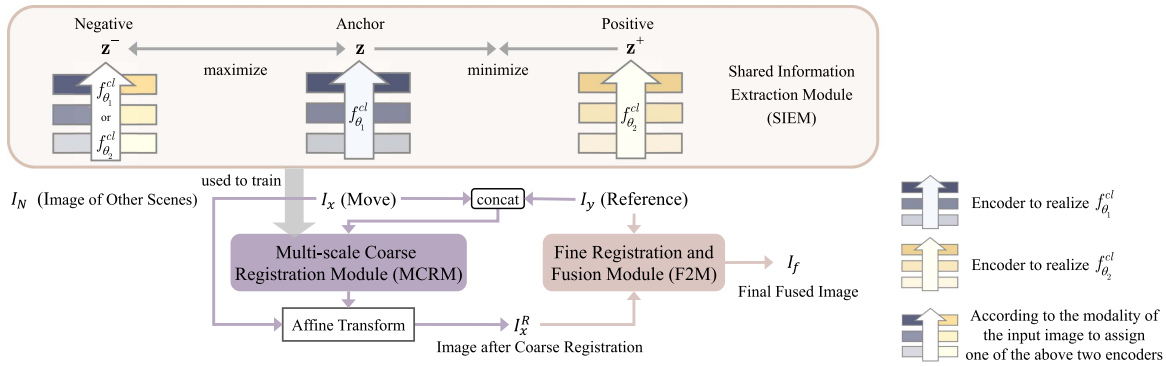


Fig. 3. Overall framework of the proposed MURF for unified multi-modal image registration and fusion (in the encoders, light to dark colors and bottom-up arrows: layers from shallow to deep).

In terms of the loss function, we design a texture loss to preserve the sharper textures in two source images and define a gradient enhancement function. Through these ways, the fused image can present more information and better visual effect.

III. Methodology

The proposed method is capable of processing multi-modal signals with offsets. It can correct the original parallaxes and generate a fused image. This section introduces the overall framework consisting of three main modules. Each module is described in detail, including the internal pipeline, loss function, network architecture, and other settings.

A. Problem Formulation

We develop a network for unified multi-modal image registration and fusion, termed as MURF. With unaligned multi-modal source images denoted as I_x and I_y, we aim to register I_x with the reference image I_y and generate the fused image I_f. To this end, the overall procedure is depicted in Fig. 3, which is divided into three main stages.

First, a shared information extraction module (SIEM) captures the information shared across multiple modalities. It helps transform the multi-modal registration challenge into mono-modal registration in the common space. The extraction functions are then used in the registration module.

Second, a multi-scale coarse registration module (MCRM) performs the global correction. The registration constraints are established using the representations extracted by SIEM and then utilized to train the networks in MCRM. MCRM outputs the image after coarse registration (I_x^R). Except for some local parallaxes where an affine model is not applicable, the images are roughly aligned.

Finally, a fine registration and fusion module (F2M) with I_x^R and I_y as input, integrates source information, and corrects local parallaxes to generate the final fused image I_f.

B. Shared Information Extraction Module (SIEM)

Each modality has its own unique attributes. For instance, RGB, NIR, and IR images represent information in different wavelength bands. CT and MRI images characterize dense structures and soft-tissue information, respectively [35]. However, some significant properties tend to be shared across modalities, e.g., objects, and geometry. The modal-independent information is useful for registration, so our goal is to map multi-modal images into some modal-independent shared space. Thus, we employ contrastive learning, where images of the same scene correspond to close representations while those of different scenes correspond to far apart representations, as shown in Fig. 3.

The multi-modal dataset consists of aligned/roughly aligned image pairs $\{I_x^i, I_y^i\}_{i=1}^K$ where K is the number of image pairs. I_x and I_y are images belonging to different modalities $\mathcal{X}$ and $\mathcal{Y}$, respectively. Their size can be represented as $H \times W \times C$ where H and W are height and width. $C=1$ for gray image or $C=3$ for RGB images. We aim to learn two functions, $f_{\theta_1}^{cl}(\cdot)$ and $f_{\theta_2}^{cl}(\cdot)$, which map images in domains $\mathcal{X}$ and $\mathcal{Y}$ to the shared latent space, with parameters θ_1 and θ_2 to be optimized, respectively. The extracted latent representations are $z_x^i = f_{\theta_1}^{cl}(I_x^i)$, $z_y^i = f_{\theta_2}^{cl}(I_y^i)$. $\{I_x^i, I_y^i\}$ are images of the same scene. Thus, $\{z_x^i, z_y^i\}$ are positive pairs and should be pulled in the latent space. $\{I_x^i, I_y^{j(j\neq i)}\}$ or $\{I_x^i, I_x^{j(j\neq i)}\}$ are multi-modal or mono-modal images of distinct scenes. Their latent representations are negative pairs and should be pushed apart.

Loss Function. Learning with complete samples imposes a tremendous strain on storage and optimization. We randomly sample a few data $\{I_x^m, I_y^m\}_{m\in\mathbf{M}}$ at a time. $\mathbf{M}$ is a subset of $\{1, 2, \ldots, K\}$ containing M indexes. Then, we tune θ_1, θ_2 with the sampled set. With a discriminating function $s\langle\cdot\rangle$ ranking high values for positive pairs while low values for negative ones, the contrastive loss function for learning $f_{\theta_1}^{cl}(\cdot)$ and $f_{\theta_2}^{cl}(\cdot)$ is defined as the InfoNCE loss [44]:

$$\mathcal{L}_c^{\mathcal{X}} = -\sum_{m\in\mathbf{M}} \log \frac{\exp(s\langle z_x^m, z_y^m\rangle \cdot \frac{1}{\tau})}{\sum\limits_{\substack{n\in\mathbf{M}\\ n\neq m}} \exp(s\langle z_x^m, z_y^n\rangle \cdot \frac{1}{\tau}) + \exp(s\langle z_x^m, z_x^n\rangle \cdot \frac{1}{\tau})}, \tag{1}$$

where τ is a temperature coefficient to adjust the dynamic range. $f_{\theta_1}^{cl}(I_x^m)$ is an anchor representation extracted from an image in domain $\mathcal{X}$. Similarly, we can anchor at domain $\mathcal{Y}$ and symmetrically construct the $\mathcal{Y}$-related contrastive loss, $\mathcal{L}_c^{\mathcal{Y}}$.

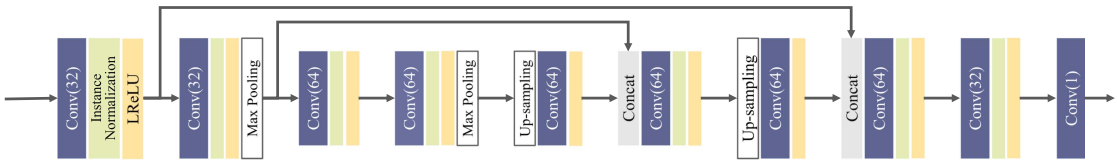


Fig. 4. Network architecture of encoders to realize $f_{\theta_1}^{cl}$ and $f_{\theta_2}^{cl}$ in the shared information extraction module (SIEM). The kernel size and stride of the convolutional layers are 3×3 and 1.

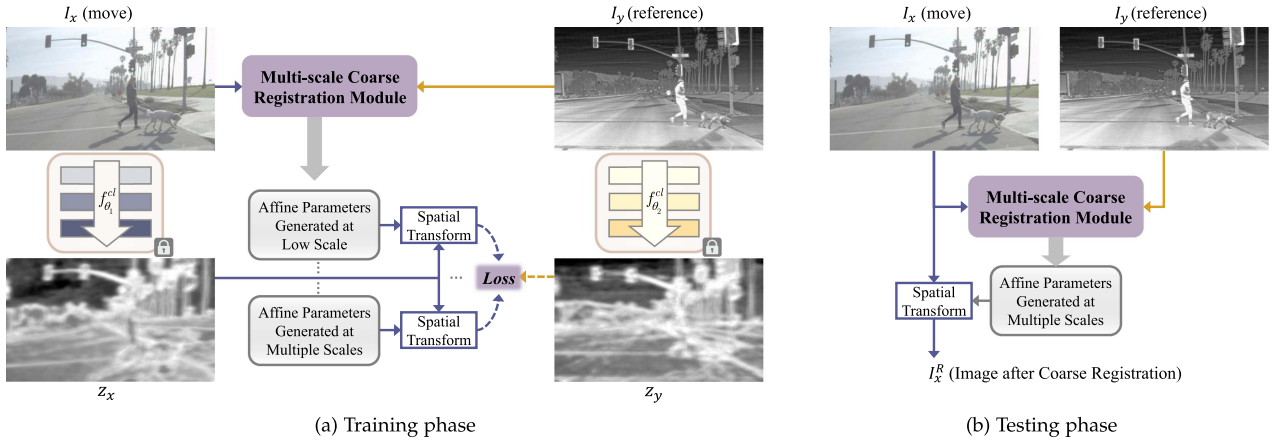


Fig. 5. Framework of the multi-scale coarse registration module (MCRM).

Thus, the contrastive loss function can be refined as:

$$\mathcal{L}_{\text{contrast}} = \mathcal{L}_c^{\mathcal{X}} + \mathcal{L}_c^{\mathcal{Y}}. \tag{2}$$

To make the extracted representations serve for registration, they should possess several characteristics: i) have the same spatial resolution as the input images; ii) provide fine structure details to ensure registration accuracy; and iii) be in a single channel to drop modal-dependent attributes. Thus, z_x and z_y are both of size $H \times W$. Considering these factors and inspired by [45], we refine the fineness of latent representations by rotational equivalence. Specifically, the implementation of $f_\theta^{cl}(\cdot)$ is replaced by $T^{-1} f_\theta^{cl}(T(\cdot))$. $T(\cdot)$ denotes the image-level rotation and $T^{-1}(\cdot)$ represents the corresponding reverse rotation in the latent space. Moreover, considering the spatial resolution of representations, the discriminating function $s\langle\cdot\rangle$ quantifies the score of a pair of representation $\{z_1, z_2\}$ with l_2 distance:

$$s\langle z_1, z_2\rangle = -\|z_1 - z_2\|_2^2. \tag{3}$$

Network Architecture. We realize the function $f_{\theta_1}^{cl}(\cdot)$ and $f_{\theta_2}^{cl}(\cdot)$ through two pseudo-siamese encoders. The network architecture of these encoders is shown in Fig. 4. The input is a source image and the output is the extracted shared latent information. It consists of ten layers and instance normalization rather than batch normalization is used as it performs a kind of style normalization [46].

C. Multi-Scale Coarse Registration Module (MCRM)

As shown in Fig. 5, MCRM takes I_x and I_y as input and generates multi-scale affine parameters for spatial transformation. In the training phase, the pre-trained $f_{\theta_1}^{cl}(\cdot)$ and $f_{\theta_2}^{cl}(\cdot)$ extract the shared information z_x and z_y. Then, MCRM is optimized by improving the registration accuracy between the deformed z_x through affine transform and z_y. In this phase, parameters in SIEM are fixed. In the testing phase, only MCRM is used to perform coarse registration.

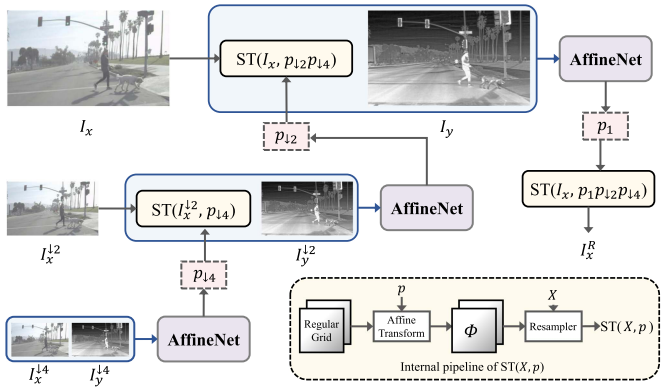


Fig. 6. Internal pipeline of MCRM to generate multi-scale affine parameters $\{p_{\downarrow 4}, p_{\downarrow 2}, p_1\}$). The implementation of spatial transform based on an image X and affine parameters p, i.e., ST(X, p), is shown in the yellow region.

Specific to MCRM, it is expected to correct long-distance global parallaxes. In single-scale networks, large kernel sizes and deep layers are necessary for wide receptive fields to capture long-distance parallaxes. To alleviate this problem, we apply a multi-scale progressive registration strategy to reduce parameter numbers and speed up convergence.

As shown in Fig. 6, the original I_x and I_y are down-sampled into lower scales, i.e., 1/2 and 1/4. After down-sampling, the long-distance parallax is more easily captured by small receptive

fields. We first use AffineNet to learn the affine parameters at the 1/4 scale, denoted as $p_{\downarrow 4} \in \mathbb{R}^{2\times 3}$. Then, we perform the rough spatial transform on $I_x^{\downarrow 2}$ at the 1/2 scale with $p_{\downarrow 4}$, represented as $\mathrm{ST}(I_x^{\downarrow 2}, p_{\downarrow 2})$. In this case, the original parallax between $I_x^{\downarrow 2}$ and $I_y^{\downarrow 2}$ is roughly reduced. On this basis, AffineNet is further applied to learn the finer affine parameters at the 1/2 scale, i.e., $p_{\downarrow 2}$. On the whole, the affine parameters for $\{I_x, I_y\}$ learned at 1/2 and 1/4 scales are represented as $p_{\downarrow 2}p_{\downarrow 4}$. Similarly, the registration process is performed at the original scale to generate the finest parameters p_1 and obtain the final output after coarse registration, i.e., $I_x^R = \mathrm{ST}(I_x, p_1 p_{\downarrow 2} p_{\downarrow 4})$.

Specific to the implementation of spatial transform, the details are provided in Fig. 6. Given an image X and the affine parameter p, we apply p on a regular sampling grid to generate a deformation field ϕ of size $H \times W \times 2$. It represents the deformation of pixels in X. The two channels of ϕ represent deviations in vertical and horizontal directions, respectively. Finally, the deformed X is resampled as:

$$\mathrm{ST}(X, p)[i + \phi_{i,j,1}, j + \phi_{i,j,2}] = X[i, j], \tag{4}$$

where i, j denote the position of the pixel.

Loss Function. The loss function of MCRM is designed as in Fig. 5. It depends on the multi-scale affine parameters $\{p_{\downarrow 4}, p_{\downarrow 2}, p_1\}$ generated in Fig. 6 and the extracted shared information $\{z_x, z_y\}$. The problem of multi-modal image registration is transformed into mono-modal registration between the deformed z_x and z_y. The affine parameters provide the deformation applied to z_x.

Therefore, we define a loss function that measures the registration accuracy of the deformed z_x and z_y to optimize the parameters of AffineNet in Fig. 6. For ease of computational tractability and for weaker sensibility to linear changes in intensity amplitudes, we use normalized cross-correlation (NCC) to measure the accuracy. The larger the NCC, the more correlated and aligned X is to Y. Thus, the loss function of MCRM is defined as:

$$\mathcal{L}_{\text{coarse}} = -\sum_{p\in\{p_{\downarrow 4}, p_{\downarrow 2}, p_1\}} \mathrm{NCC}(\mathrm{ST}(z_x, p), z_y). \tag{5}$$

Before computing NCC, the deformed z_x and z_y are normalized through the following formulation:

$$z_x^* = \mathrm{clip}\left(\frac{z_x - z_{\min}}{z_{\max} - z_{\min}}\right), \quad z_y^* = \mathrm{clip}\left(\frac{z_y - z_{\min}}{z_{\max} - z_{\min}}\right), \tag{6}$$

where $z_{\min}$ is the minimum of $\min(z_x)$ and $\min(z_y)$. $z_{\max}$ is the maximum of $\max(z_x)$ and $\max(z_y)$. $\mathrm{clip}(\cdot)$ represents clipping to [0, 1]. Then, NCC is defined as:

$$\mathrm{NCC}(X, Y) = \frac{\sum_{i=1}^{W}\sum_{j=1}^{H}(X_{i,j} - \bar{X})(Y_{i,j} - \bar{Y})}{\sqrt{\sum_{i=1}^{W}\sum_{j=1}^{H}(X_{i,j} - \bar{X})^2}\sqrt{\sum_{i=1}^{W}\sum_{j=1}^{H}(Y_{i,j} - \bar{Y})^2}}, \tag{7}$$

where X, Y are two images. $\bar{X}, \bar{Y}$ denote their mean values.

Network Architecture. MCRM employs three AffineNets to generate the affine parameters. The AffineNets are pseudo-siamese networks and shown in Fig. 7. Deformable convolution layers are applied as they can augment the traditional regular receptive fields with horizontal and vertical offsets, which are learned from additional convolution layers. The deformable convolution layers refer to deformations in unaligned images for higher registration accuracy and stronger robustness. We apply many layers, large kernel sizes, and max pooling layers to capture deformations. Then, the feature maps are mapped into a 128-dimensional vector with the global average pooling (GAP) layer. In the training phase, we apply dropout to further improve the performance. Finally, a 128-dimensional vector is fed into a fully connected layer to generate a 6-dimensional parameter and reshaped into size 2×3 as the output affine parameters.

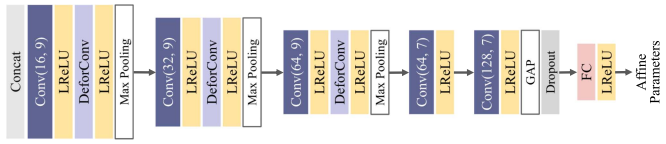


Fig. 7. Network architecture of AffineNet. "Conv(c, k)": the channel of output features is c and the kernel size is $k \times k$; "GAP": global average pooling; "FC": fully connected layer; "DeforConv": deformable convolution layer. The stride of convolution layers is 1.

D. Fine Registration and Fusion Module (F2M)

F2M realizes image fusion and corrects the local non-rigid parallaxes to generate the final aligned and fused image. The framework of F2M is shown in Fig. 8. In terms of the generation process, I_x^R and I_y are first fed into the deformation block and spatial transform to correct local parallaxes. The output of the spatial transform is the deformed I_x^R, denoted as I_x^F. Then, I_x^F and I_y are fused through the subsequent extraction layers, gradient channel attention block, and reconstruction layers for image fusion.

The training process is split into two phases. As the fused image should provide feedback for the fine registration, F2M realizes image fusion in the first phase. We optimize the fusion-related parameters and the deformation block is excluded. The deformation block depends on initialized parameters to generate the deformation field, which automatically tends to be identical. Then, I_f nearly combines the scene information of I_x^R and I_y, and renders their parallaxes in a single image. In the second phase, F2M realizes the fine registration based on I_f and the inverse deformation field. The fusion-related parameters which have been optimized are fixed and the deformation block is optimized.

1) Image Fusion: Image fusion is realized in fusion-related layers besides the deformation block and spatial transform. As the fused image is expected to present a large amount of information about scenes, we design the fusion-related part in terms of loss function and network architecture to present clear and rich textures in the fused image. For color images such as RGB and PET images, we transform them into YCbCr space and fuse the luminance information (Y channel) with the other source image. Then, the fused image is concatenated with chrominance (Cb and Cr channels) and transformed into RGB space to generate the final fused RGB image.

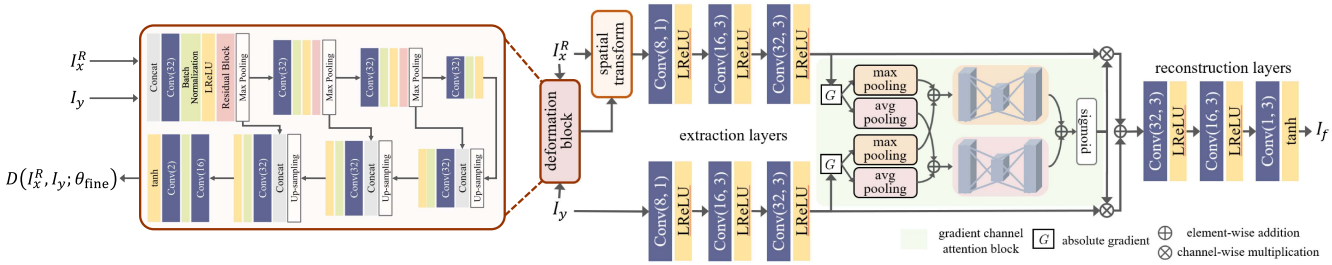


Fig. 8. Framework of the fine registration and fusion module (F2M).

Loss Function. As the fused image should retain the information from source images, we define a similarity loss $\mathcal{L}_{\text{sim}}$ to constrain the similarity between the fused image and source images. Besides, some textures may suffer from poor visibility due to low illumination, improper correction, or other factors. If they are enhanced in the fused image, it will further improve the visual effect. Thus, we define a texture loss $\mathcal{L}_{\text{texture}}$ to preserve the salient textures and enhance the textures with poor visibility. With a hyper-parameter δ controlling the trade-off, the fusion loss is defined as:

$$\mathcal{L}_{\text{fuse}} = \mathcal{L}_{\text{sim}} + \delta \mathcal{L}_{\text{texture}}. \tag{8}$$

$\mathcal{L}_{\text{sim}}$ constrains the structural similarity according to light, contrast, and structure as defined by the structural similarity index measure (SSIM) [47]. It is denoted as:

$$\mathcal{L}_{\text{sim}} = 1 - \frac{1}{2}\text{SSIM}(I_f, I_x^F) - \frac{1}{2}\text{SSIM}(I_f, I_y). \tag{9}$$

For texture preservation and enhancement, we first generate a binary gradient mask by comparing the absolute gradients of the same location in I_x^F and I_y, defined as:

$$M[i,j] = \begin{cases} 1, & \left|\nabla I_x^F[i,j]\right| > \left|\nabla I_y[i,j]\right|, \\ 0, & \text{else}. \end{cases} \tag{10}$$

This mask is used in the texture loss to preserve the larger gradients. Thus, the texture loss is defined as:

$$\begin{aligned} &\mathcal{L}_{\text{texture}} \\ &= \frac{1}{HW}\sum_{i=1}^{W}\sum_{j=1}^{H} M[i,j] \\ &\cdot \left(\nabla I_f[i,j] - \frac{\nabla I_x^F[i,j]}{|\nabla I_x^F[i,j]|}\left|\nabla I_x^F[i,j]\right|^{\gamma}\right) \\ &+(1-M[i,j]) \cdot \left(\nabla I_f[i,j] - \frac{\nabla I_y[i,j]}{|\nabla I_y[i,j]|}\left|\nabla I_y[i,j]\right|^{\gamma}\right), \end{aligned} \tag{11}$$

where $\frac{I[i,j]}{|I[i,j]|}$ is used to keep the sign of gradient. Similar to the gamma correction, we use the power function to enhance gradients. γ is set to 0.7 in this work.

Network Architecture. As shown in Fig. 8, the network architecture for image fusion consists of extraction layers, gradient channel attention block for texture preservation, and reconstruction layers. We aggregate the absolute gradients as they are a better representation of information richness in feature maps. The information is aggregated by jointly using max-pooling and average-pooling. Then, the two-branch results are added and fed into two individual multi-layer perceptrons to generate shared channel-wise attention weights. Then, the reconstruction layers map the additive features back to the image domain to generate I_f.

Fig. 9. Generation and implementation of a non-rigid deformation field. According to the solved inverse deformation field, we can inverse the deformed image back. The mean absolute error between the first and last images is 2.141 (mainly caused by areas that cannot be recovered around).

2) Fine Registration: The deformation block takes I_x^R and I_y as input and generates the deformation field to correct local parallaxes through spatial transformation. To train the block, we artificially create a locally smooth non-rigid deformation field ϕ^{nr}, as shown in Fig. 9. ϕ^{nr} is applied to an aligned/roughly aligned image in the domain $\mathcal{X}$ to create a deformed image. The optimization of the block relies on two aspects. First, the artificially set deformation field theoretically corresponds to an inverse deformation ϕ^{inv}. It can inversely transform the deformed image into the original appearance and can be used for supervision. However, the image pairs in some publicly available datasets are not strictly aligned and still suffer from some small parallaxes. Thus, the inversion deformations are not completely accurate. Second, we also rely on the characteristics of the fused image for correction. It is easy to observe that any misalignments in I_f will decrease the gradient sparsity. We encourage the sparsity of ∇I_f and penalize the salient gradients that should be corrected.

Loss Function. The loss function of the deformation block consists of two terms. The generated deformation should be roughly similar to ϕ^{inv}, and the gradient sparsity of I_f should be encouraged. The loss function is defined as:

$$\mathcal{L}_{\text{fine}} = \|D(I_x^R, I_y; \theta_{\text{fine}}) - \phi^{\text{inv}}\|_2 + \eta \sum_{i=1}^{W}\sum_{j=1}^{H} \psi_0(I_f[i,j]), \tag{12}$$

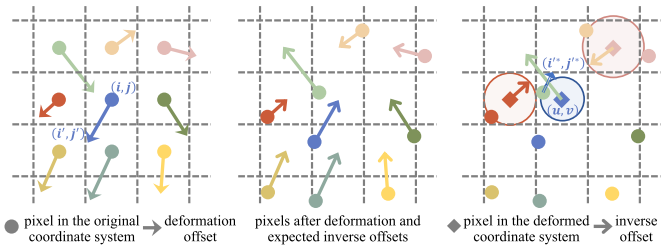


Fig. 10. Illustration of generating the inverse deformation.

where $D(\cdot)$ denotes the deformation generated by the deformation block. θ_{fine} denotes the parameters in this block. η is a hyper-parameter. Inspired by [48], we define a sparse loss function $\psi_0(\cdot)$ to effectively approximate L_0 sparsity:

$$\psi_0(I_f[i,j]) = \begin{cases} \frac{1}{\epsilon^2}\left|\nabla I_f[i,j]\right|^2, & \left|\nabla I_f[i,j]\right| < \epsilon, \\ 1, & \text{else.} \end{cases} \tag{13}$$

When $\left|\nabla I_f[i,j]\right| < \epsilon$, $\psi(\cdot)$ is continuous, a necessary condition to form a loss function. Finally, $D(I_x^R, I_y; \theta_{\text{fine}})$ is applied on I_x^R to generate the fine-registered image I_x^F.

Network Architecture. The network architecture of the deformation block is shown in Fig. 8. It adopts the form of U-Net. Max pooling and deep layers are used to expand receptive fields to capture corresponding pixels after deformation. Residual blocks are also applied to compare and learn the offsets. The final output of this block is a deformation field of size $H \times W \times 2$ with the two channels representing their horizontal and vertical offsets respectively.

Generation of Inverse Deformation. We discuss the generation of the inverse deformation field. For a deformation field ϕ^{nr}, the existing methods may intuitively and directly set the inverse deformation field as $-\phi^{nr}$. However, as the deformed coordinate systems has been warped and distorted compared with the original coordinate system, $-\phi^{nr}$ is not the correct solution. In this part, we analyze the relationship between the original and deformed coordinate systems and aim to solve for a more accurate solution for the inverse deformation field.

In the original coordinate system, each pixel corresponds to its horizontal and vertical offsets, as shown in the first image in Fig. 10. After deformation, its position is:

$$i' = i + \phi^{\text{nr}}_{i,j,1}, \quad j' = j + \phi^{\text{nr}}_{i,j,2}, \tag{14}$$

where $\{i, j\}$ indicates the original coordinates of a pixel with $i \in \{1, \ldots, W\}, j \in \{1, \ldots, H\}$. $\{i', j'\}$ represents the deformed position. In the deformed image, if $\{-\phi^{\text{nr}}_{i,j,1}, -\phi^{\text{nr}}_{i,j,2}\}$ are applied to $\{i', j'\}$, this pixel can be transformed into its original position, as shown in the second image in Fig. 10. The inverse deformation field ϕ^{inv} can be ideally set as:

$$\phi^{\text{inv}}_{i',j',1} = -\phi^{\text{nr}}_{i,j,1}, \quad \phi^{\text{inv}}_{i',j',2} = -\phi^{\text{nr}}_{i,j,2}. \tag{15}$$

However, the deformed coordinate system is different from the original one as the pixels are randomly scattered. The pixel in the deformed coordinate system may not correspond to the pixel in the same position in the original coordinate system as they are decimals. In this case, for a pixel $\{u, v\}$ in the deformed coordinate system, we look for the closest deformed point and rely on its offset to set the inverse offset. For example, in the third image of Fig. 10, for the blue pixel (diamond) in the deformed coordinate system, the closest deformed point is the green one (circle). The offset of the blue pixel is set as the inverse offset of the green point. Mathematically, for a pixel $\{u, v\}$ in the deformed coordinate system, the closest deformed point $\{i'^*, j'^*\}$ is:

$$(i'^*, j'^*) = \underset{\{i',j'\} \in \mathcal{R}_{u,v}}{\arg\min} \sqrt{(u - i')^2 + (v - j')^2}, \tag{16}$$

where $\mathcal{R}_{u,v}$ denotes the neighborhood near $\{u, v\}$ to narrow the solution space and improve efficiency. As $\{i'^*, j'^*\}$ is deformed from $\{i^*, j^*\}$ in the original coordinate system, the inverse deformation field ϕ^{inv} can be solved as:

$$\phi^{\text{inv}}_{u,v,1} = -\phi^{\text{nr}}_{i^*,j^*,1}, \quad \phi^{\text{inv}}_{u,v,2} = -\phi^{\text{nr}}_{i^*,j^*,2}. \tag{17}$$

Fig. 10 provides the generated inverse deformation field and the inversely deformed image obtained as:

$$I^{\text{inv}}[u + \phi^{\text{inv}}_{u,v,1}, v + \phi^{\text{inv}}_{u,v,2}] = I'[u, v], \tag{18}$$

where I' is the deformed image transformed from the original image I. The slight differences between I^{inv} and I (reported in the caption of Fig. 9) proves its correctness.

IV. Experiments and Results

We compare MURF with both state-of-the-art (SOTA) multi-modal image registration and fusion methods. Experiments are implemented on four types of multi-modal data (four tasks) to verify generalization. At the end of subsections of registration and fusion, the ablation study is performed to verify effectiveness of some designs and settings.

A. Implementation Details

Multi-Modal Images. We test the proposed method on a variety of multi-modal data, including i) RGB and infrared (RGB-IR) images; ii) RGB and near-infrared (RGB-NIR) images; iii) positron emission tomography and magnetic resonance imaging (PET-MRI) images; and iv) computed tomography and MRI (CT-MRI) images. The data comes from publicly accessible datasets, including *RoadScene*[1] (RGB-IR images), *VIS-NIR Scene*[2] (RGB-NIR images), and *Harvard*[3] (PET-MRI and CT-MRI images). As images in these datasets are aligned/roughly aligned, we manually build deformations to obtain unaligned images for training and testing. The deformations are applied to RGB, PET, and CT images (three specific types of I_x described in Section III-A). The reference images I_y are IR, NIR, and MRI images, respectively.

Training Details. We select images from the above datasets and apply deformations. Then, the unaligned images are cropped into patches as the training data. For each task of each module,

[1]https://github.com/hanna-xu/RoadScene
[2]http://matthewalunbrown.com/nirscene/nirscene.html
[3]http://www.med.harvard.edu/AANLIB/home.html

TABLE I
TRAINING DETAILS OF MODULES ON FOUR TASKS. THE LEARNING RATE IS SET IN TWO WAYS: FOR SIEM AND MCRM, IT IS INITIALLY SET TO 0.0002 WHICH DECAYS EXPONENTIALLY; FOR F2M, IT IS SET TO A FIXED VALUE OF 0.0001. IN F2M, THE TRAINING SETTINGS OF THE DEFORMATION BLOCK AND FUSION LAYERS ARE THE SAME

Settings		Patch Size	Batch Size	Epoch	Learning Rate
SIEM	RGB-IR	256×256	32	50	0.0002
	RGB-NIR	256×256	32	50	0.0002
	PET-MRI	256×256	32	50	0.0002
	CT-MRI	256×256	32	50	0.0002
MCRM	RGB-IR	256×256	32	100	0.0002
	RGB-NIR	256×256	16	100	0.0002
	PET-MRI	256×256	32	100	0.0002
	CT-MRI	256×256	32	100	0.0002
F2M	RGB-IR	512×512	32	15	0.0001
	RGB-NIR	256×256	8	20	0.0001
	PET-MRI	256×256	32	50	0.0001
	CT-MRI	256×256	32	50	0.0001

Algorithm 1: Overall Description of MURF.

1: **Training phase:**
2: Update the parameters θ_1, θ_2 in SIEM by minimizing $\mathcal{L}_{\text{contrast}}$ defined in (2) to learn $f^{cl}_{\theta_1}(\cdot), f^{cl}_{\theta_2}(\cdot)$;
3: Fix θ_1, θ_2, and use $f^{cl}_{\theta_1}(\cdot), f^{cl}_{\theta_2}(\cdot)$ to update parameters in MCRM by minimizing $\mathcal{L}_{\text{coarse}}$ in (5);
4: Fix the parameters in MCRM;
5: Fix θ_{fine} and update the fusion-related parameters in F2M by minimizing $\mathcal{L}_{\text{fuse}}$ defined in (8);
6: Update parameters in the deformation block of F2M, i.e., θ_{fine}, by minimizing $\mathcal{L}_{\text{fine}}$ defined in (12);
7: **Testing phase:**
8: Feed the unaligned multi-modal source images $\{I_x, I_y\}$ into MURF, and use SIEM, MCRM, and F2M to generate the aligned fusion result I_f.

the settings of patch size, batch size, epoch, learning rate, *etc*., are provided in Table I and the process of training the three modules is summarized as Algorithm 1. All the tasks use the Adam optimizer. The registration-related modules/blocks are first trained on small patches and then fine-tuned on large-resolution images. The hyper-parameters are set as: $\tau = 1, \delta = 10, \eta = 0.001, \epsilon = 0.1$. Experiments are performed on NVIDIA Geforce GTX Titan X GPU and 2.4 GHz Intel Core i5-1135G7 CPU.

B. Shared Information Extraction

We compare shared information extraction module (SIEM) with some existing methods, including WLD [10], NTG [49], and SCB [12]. The shared information z_x, z_y extracted from I_x, I_y by different methods are shown in Fig. 11. The effectiveness and generalization are validated on four tasks. There are significant intensity and structure differences in multi-modal images. Among the results, our results are the most consistent in visual effect. To compare differences clearly, we scan the pixel intensities of a column (RGB images are first transformed into gray images). The scanned results objectively prove that our SIEM can effectively decreases modal disparities of images and capture their shared information. Notably, when processing multi-modal medical images with larger modal variances, the comparison methods almost cease to be effective as the extracted information shows conspicuous differences. It further validates the effectiveness and generalization of our method.

TABLE II
COMPARISON OF NORMALIZED CROSS-CORRELATION (NCC) BEFORE AND AFTER SHARED INFORMATION EXTRACTION WITH DIFFERENT METHODS (MEAN AND STANDARD DEVIATION ARE SHOWN, *BOLD*: OPTIMAL, UNDERLINE: SUBOPTIMAL)

	RGB-IR	RGB-NIR	PET-MRI	CT-MRI
images	-0.120±0.445	0.461±0.229	0.444±0.131	0.459±0.133
WLD [10]	0.212±0.106	0.366±0.238	0.090±0.019	0.114±0.050
NTG [49]	0.313±0.088	0.518±0.167	0.084±0.043	0.245±0.066
SCB [12]	0.437±0.100	0.628±0.159	0.326±0.078	0.556±0.093
SIEM (ours)	**0.702±0.182**	**0.847±0.083**	**0.979±0.058**	**0.946±0.073**

For quantitative comparison, we use NCC to quantify the correlations between I_x, I_y and z_x, z_y, respectively. In each task, quantitative results are tested on 180 aligned/roughly aligned pairs. Similarly, RGB images are transformed to gray images when calculating. Quantitative results before and after extraction are reported in Table II. Horizontally, RGB-IR shows a substantially lower correlation than other multi-modal data, indicating the most significant modal variance. The PET, CT, and MRI images also look modally different while the background (dark) regions pull up the correlation. Vertically, SIEM improves the correlation substantially, especially for RGB-IR. Compared with other methods, our results also show significant improvements. It can be indicated that z_x and z_y extracted by our SIEM are in a common space. Thus, we can use the learned $f^{cl}_{\theta_1}(\cdot)$ and $f^{cl}_{\theta_2}(\cdot)$ to guide the training of the unsupervised multi-scale coarse registration module.

C. Multi-Modal Image Registration

1) Qualitative Results: We compare our multi-scale coarse registration module (MCRM) with some SOTA registration methods, including SIFT [50], DASC [51], [52], NTG [49], SCB [12], and MIDIR [53]. Among these methods, SIFT, NTG, and SCB are only capable of dealing with rigid deformations as they estimate the affine parameters. DASC and MIDIR can deal with non-rigid deformation as they estimate flow fields. Qualitative registration results on four tasks are shown in Fig. 12. In each group, the deformed image and the reference source image are superimposed to exhibit misalignments. Their gradients are also superimposed for auxiliary comparison. Especially, it is difficult to subjectively distinguish whether the structures in medical images are corresponding. Thus, we mark five pairs of identically positioned points on the original aligned PET-MRI and CT-MRI image pairs. Then, the unregistered images are artificially created. The registration accuracy of medical images can be observed through the distances between marked points. The overlapped points indicate the correct deformation.

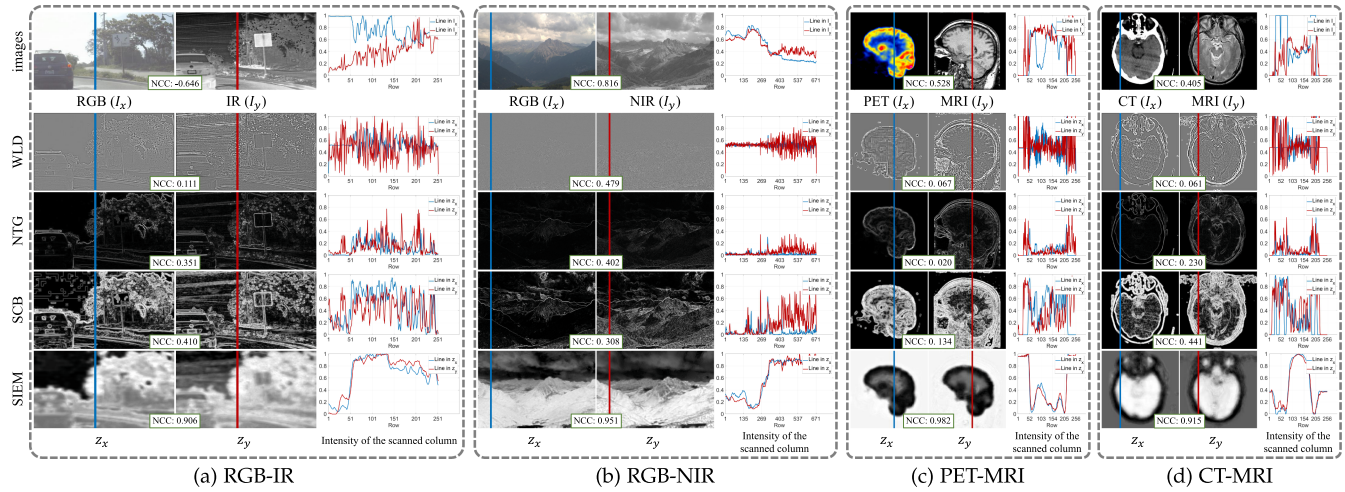


Fig. 11. Qualitative comparison of extracted shared information z_x, z_y from multi-modal images I_x, I_y. The normalized cross-correlation (NCC) between paired data is reported where paired data is spliced (larger value: higher correlation).

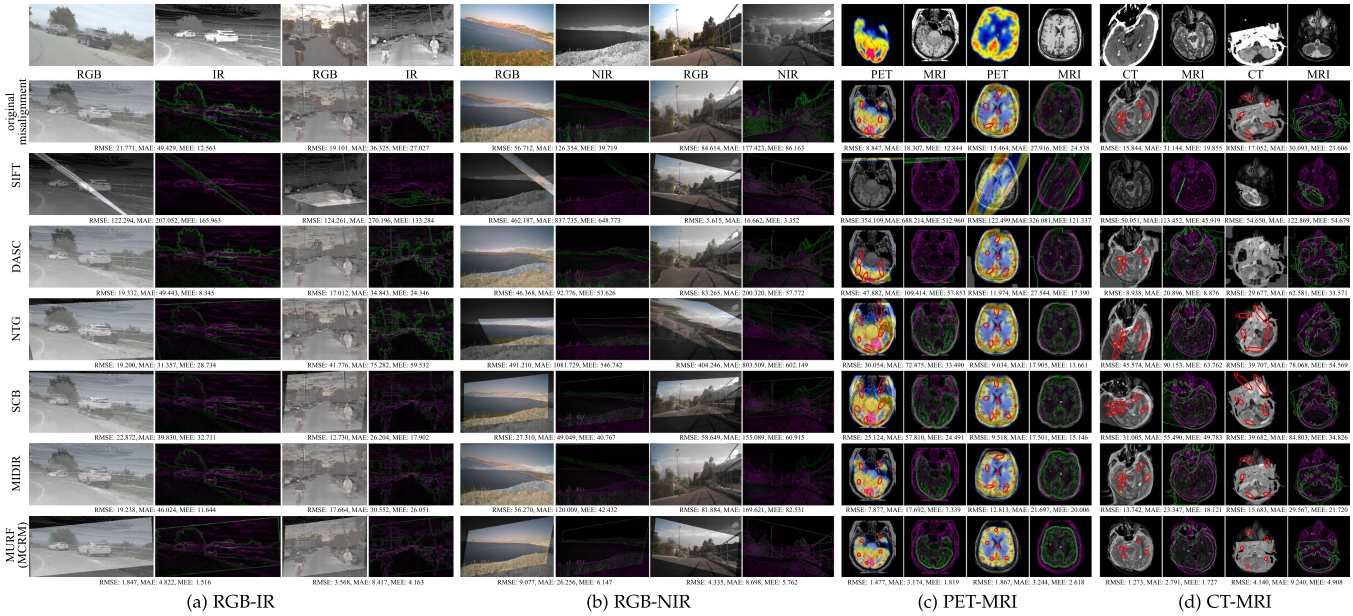


Fig. 12. Qualitative registration results on multi-modal image pairs. In each group, the first column: superimposed images (original misalignment: source images; other rows: deformed image and the reference source image); the second column: their superimposed gradients. The marked points in PET images are green square points and those in CT/MRI images are black/white square points. In superimposed images, the two points that should be in the same position are circled. RMSE, MAE, and MEE (introduced in Section IV-C2) measure the accuracy (smaller value: higher accuracy).

As shown in Fig. 12, SIFT fails to align RGB-IR, PET-MRI, and CT-MRI images and even aggravates the deformation. The same phenomenon occurs when NTG aligns RGB-NIR and PET-MRI images. However, SIFT can successfully deal with some RGB-NIR scenarios, such as the last group in Fig. 12(b). It corrects most deformations efficiently except for some slight offsets at the rightmost railing. Similarly, NTG can alleviate some offsets in RGB-IR and CT-MRI images, such as the first group in Fig. 12(a) and some regions in Fig. 12(d). These methods are applicable to specific types of data but suffer poor generalization. By comparison, DASC universally suffers severe geometric distortion in all image pairs, which is easily observed in Fig. 12(c) and (d) and the last group in Fig. 12(b). SCB and MIDIR exhibit the tendency to narrow the original offsets, while their registration accuracy is still inferior to our MCRM.

These results demonstrate that our MCRM outperforms the SOTA methods with higher registration accuracy and better generalization for multiple tasks including RGB-IR, RGB-NIR, PET-MRI, and CT-MRI image registration.

2) Quantitative Results: As illustrated in Fig. 13, five pairs of point landmarks are artificially labelled in each image pair and scattered throughout the image. The source points in I_x are $\{(w_s^k, h_s^k)\}_{k=1}^5$ and the target ones in I_y are $\{(w_t^k, h_t^k)\}_{k=1}^5$. In the deformed image, the source points are transformed

TABLE III
QUANTITATIVE REGISTRATION PERFORMANCE COMPARISON RESULTS OF THE MULTI-SCALE COARSE REGISTRATION MODULE (MCRM) IN THE PROPOSED MURF AND FIVE COMPETITORS (*BOLD*: OPTIMAL, UNDERLINE: SUBOPTIMAL)

(a) RGB-IR

Metrics	RMSE↓	MAE↓	MEE↓
Original RGB	12.843±5.896	27.405±13.502	16.168±9.403
SIFT [50]	907.236±4789.188	2008.480±10455.677	639.299±2757.836
DASC [51], [52]	14.114±14.730	33.268±31.672	14.678±22.059
NTG [49]	6.003±11.262	13.209±23.889	6.930±14.541
SCB [12]	7.884±7.986	17.319±17.195	10.158±11.326
MIDIR [53]	12.221±5.764	26.485±12.873	14.702±9.528
MCRM	**3.114±4.380**	**7.013±10.310**	**3.372±5.123**

(b) RGB-NIR

Metrics	RMSE↓	MAE↓	MEE↓
Original RGB	38.642±51.443	87.832±160.558	41.548±24.168
SIFT [50]	20.466±145.695	49.095±341.505	22.185±167.694
DASC [51], [52]	32.091±54.134	81.759±166.256	25.053±27.627
NTG [49]	83.812±227.921	193.080±547.134	83.623±232.021
SCB [12]	21.315±52.753	51.719±159.197	19.199±28.070
MIDIR [53]	37.923±51.494	87.333±160.682	40.057±23.832
MCRM	**14.630±50.727**	**37.156±156.790**	**10.651±16.889**

(c) PET-MRI

Metrics	RMSE↓	MAE↓	MEE↓
Original PET	8.388±4.595	17.517±9.447	10.351±6.350
SIFT [50]	141.160±303.922	313.892±751.092	160.153±274.391
DASC [51], [52]	14.980±9.555	32.001±22.211	17.983±10.648
NTG [49]	16.892±8.221	35.981±20.727	21.008±9.408
SCB [12]	13.215±3.806	27.752±9.734	16.258±4.354
MIDIR [53]	6.758±3.176	13.931±6.626	8.576±4.688
MCRM	**2.017±1.846**	**4.027±4.003**	**2.589±2.388**

(d) CT-MRI

Metrics	RMSE↓	MAE↓	MEE↓
Original CT	10.507±4.585	22.040±9.459	12.573±6.635
SIFT [50]	159.342±550.683	349.084±1308.813	183.043±493.447
DASC [51], [52]	9.293±8.063	22.082±22.023	9.197±7.651
NTG [49]	6.742±10.945	14.073±21.689	8.602±14.861
SCB [12]	7.595±9.064	16.362±19.349	9.503±11.496
MIDIR [53]	7.996±3.291	16.938±7.482	9.938±4.433
MCRM	**2.387±1.292**	**4.949±2.644**	**3.083±1.849**

Fig. 13. Illustration of point landmarks. Points represented as "o" in I_x are the source points $\{(w_s^k, h_s^k)\}_{k=1}^5$ and those represented as "+" in I_y are the target points $\{(w_t^k, h_t^k)\}_{k=1}^5$.

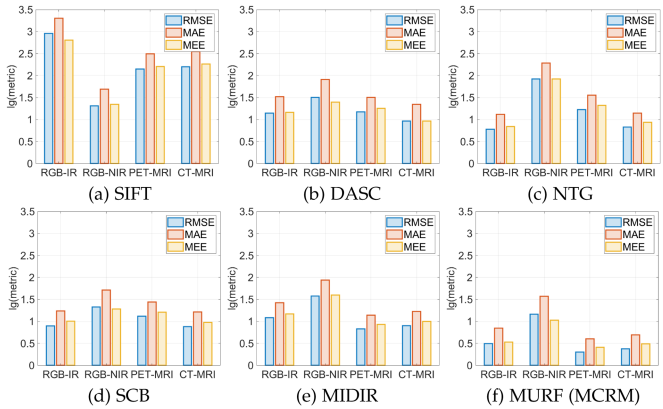


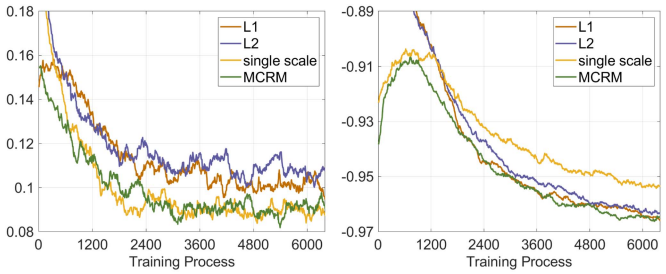
Fig. 14. Quantitative results of each registration method on the four tasks are concentrated in each subfigure to verify the generalization (results are displayed on a log scale).

into $\{(w_r^k, h_r^k)\}_{k=1}^5$, which are expected to approach the target points. The registration accuracy is evaluated with the euclidean distances between $\{(w_r^k, h_r^k)\}_{k=1}^5$ and $\{(w_t^k, h_t^k)\}_{k=1}^5$. We compare the distances through root mean square error (RMSE), max square error (MAE), and median square error (MEE).

The quantitative results are tested on 50 RGB-IR, 155 RGB-NIR, 100 PET-MRI, and 100 CT-MRI image pairs respectively. As reported in Table III, our multi-scale coarse registration module (MCRM) achieves the optimal performances on three metrics for four tasks. The smallest standard deviation of our MCRM in each task also demonstrates its universality and stability. Some competitors show small mean but large standard deviation. It demonstrates that they perform well in some scenarios while not in others.

In terms of specific tasks, NTG achieves the suboptimal performances on RGB-IR and CT-MRI image registration and MIDIR achieves the suboptimal performance on PET-MRI image pairs. SIFT shows the suboptimal registration performance on RGB-NIR image pairs while showing poor performances on other tasks. The reason is that the performance of SIFT descriptor is still limited against modal variance. Table II reports that RGB-NIR images have the highest correlation (background in medical images pull up their correlation), so SIFT can register RGB-NIR images well, but fails on other multi-modal images with low correlation. Therefore, we further compare the generalization of different methods. The quantitative results of each method on the four tasks are concentrated in each subfigure in Fig. 14. In line with the above analysis, SIFT outperforms other competitors on RGN-NIR but fails on other tasks, indicating poor generalization. As for other methods, their performances on RGB-NIR images are generally inferior to the performances on other image pairs. The reason is that the spatial resolution of RGB and NIR images is significantly higher than those of other types of images, resulting in relatively high euclidean distances between point sets. In Fig. 14(c), the inapplicability of NTG to RGB-NIR images further exacerbates its performance differences across tasks. Overall, our method shows generally optimal performance on the four tasks, rather than only applicable to specific data.

3) Ablation Study: We validate the effectiveness of several elements in the multi-scale coarse registration module (MCRM), including the shared information extraction module (SIEM) to mitigate the multi-modal challenge, the multi-scale progressive strategy, and the registration loss function. The ablation study



(a) $-\mathrm{NCC}(I_x^R, I_y)$ (b) $-\mathrm{NCC}(z_x^R, z_y)$

Fig. 15. Ablation results during the training process where the registration performances are evaluated by the normalized cross correlation both in the level of multi-modal image and extracted information in the common space. Smaller values indicate better registration performance.

Fig. 16. Pipeline of single-scale coarse registration module.

is performed on the representative RGB-IR image registration because RGB-IR images suffer from the lowest correlation in all multi-modal combinations as described in Table II. As distinct experimental settings will result in different loss functions, we evaluate the registration accuracy under different settings in a unified manner. We evaluate the accuracy from two aspects, including the accuracy of multi-modal images, i.e., $-\mathrm{NCC}(I_x^R, I_y)$, and the accuracy of the extracted common information, i.e., $-\mathrm{NCC}(z_x^R, z_y)$.

SIEM to Mitigate Multi-Modal Challenge. To evaluate the effectiveness of $z_x = f_{\theta_1}^{cl}(I_x)$, $z_y = f_{\theta_2}^{cl}(I_y)$ for multi-modal image registration, we replace the registration loss function in (5). Alternatively, we replace the registration loss with the multi-modal images I_x, I_y/results of SCB transform [12]. These approaches both encounter the gradient explosion when training, indicating that they are difficult for network optimization. The shared information extracted by SIEM is applicable for optimization, proving the effectiveness of SIEM in mitigating modal variances.

From the other point of view, the slight differences of NCC, L1, and L2 losses on registration performance in Fig. 15 also demonstrate the effectiveness of SIEM. In (5), the registration loss is based on z_x and z_y, which are expected to be in a common space. NCC shows weak sensibility to intensity differences while L1 and L2 losses do not. If z_x and z_y show significant modal variance, it will be difficult to use L1/L2 loss to optimize. However, in Fig. 15, the applications of NCC, L1, and L2 losses show slight differences. It demonstrates that the common space learned by SIEM mitigates original modal variances.

Multi-Scale Progressive Strategy. To validate the effectiveness of multi-scale strategy, we change it with the single-scale strategy in Fig. 16. Also, we redefine the registration loss in (5) in a single-scale form, represented as:

$$\mathcal{L}_{\text{coarse}} = -\mathrm{NCC}(\mathrm{ST}(z_x, p_1), z_y). \tag{19}$$

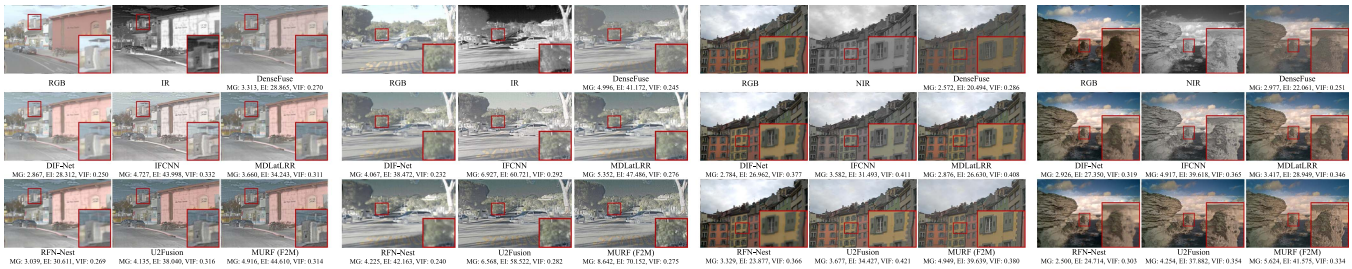
Under the single-scale strategy, the registration loss in the training process is also shown in Fig. 15. Although the differences between multi-scale and single-scale strategies are small in $-\mathrm{NCC}(I_x^R, I_y)$, they still show significant performance disparities in $-\mathrm{NCC}(z_x^R, z_y)$. Comprehensively, the multi-scale strategy is superior to the single-scale strategy.

Registration Loss Function. In MCRM, for ease of computational tractability and weaker sensibility to linear intensity changes, we use NCC to evaluate the registration accuracy. To validate its effectiveness, we replace NCC in (5) with L1 and L2 losses, respectively. As shown in Fig. 15(a), NCC achieves better registration performance than L1 and L2 losses in the image level. In Fig. 15(b), although NCC, L1, and L2 losses finally fall to the same loss range in the level of extracted information, the application of NCC loss stills shows the fastest convergence speed.

D. Multi-Modal Image Fusion

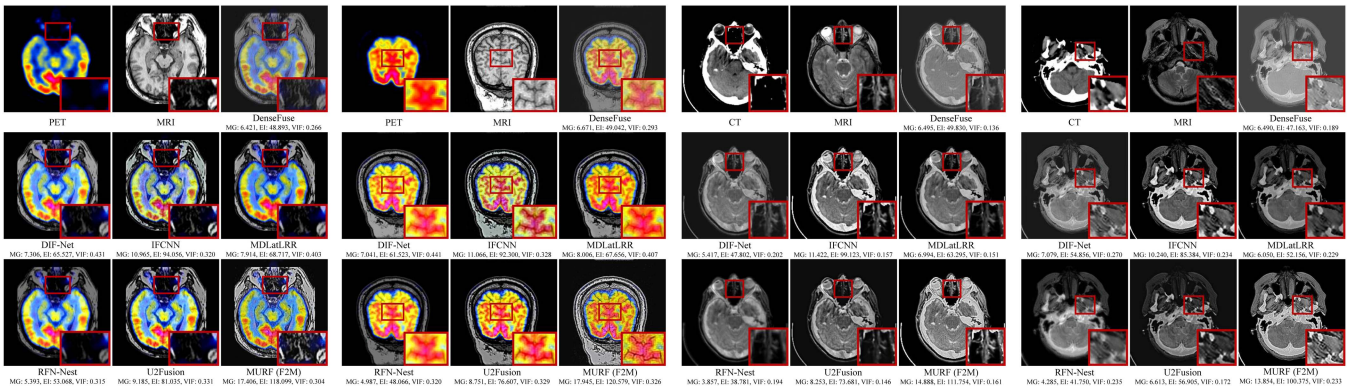
1) Qualitative Results: We compare F2M with some SOTA fusion methods, including DenseFuse [54], DIF-Net [55], MDLatLRR [56], IFCNN [57], RFN-Nest [33], and U2Fusion [6]. We consider the condition where source images contain local non-rigid parallaxes. We compare the abilities of different methods to handle local offsets and fusion performances. The qualitative results of RGB-IR and RGB-NIR images are shown in Fig. 17. In Fig. 17(a) and (c), the source images suffer from obvious non-rigid parallaxes. The parallaxes remain in the results of competitors, resulting in overlapping shadows, blurred textures, or confusing scene descriptions. In our result, the parallaxes are adjusted to provide a clearer scene description. Moreover, in Fig. 17(b) and (d), the source images are almost aligned, where we focus on comparing fusion performances. Among the competitors, IFCNN, MDLatLRR, and U2Fusion achieve sharper textures. Significant color distortions exist in DenseFuse and IFCNN. Our results exhibit clear appearance and little color distortion. Moreover, with the gradient enhancement, our results enhance the original textures rather than just preserving them.

The qualitative results of medical images are shown in Fig. 18. Due to the large modal differences between medical modalities, few textures in PET/CT images, and high freedom degree of non-rigid transformation, it is hard to generate accurate inverse deformation fields. As the image pairs in the *Harvard* dataset are aligned, we directly compare the fusion performances of F2M and other fusion methods. F2M shows three advantages. First, our results are not interfered by useless background information. In Fig. 18(a) and (c), the background in PET/CT images provides little information. The introduction of useless information in competitors leads to the information distortion of MRI images while our results preserve the information in MRI images. Second, F2M preserves the source information in balance. In Fig. 18(b) and (d), the competitors excessively preserve the functional information in PET images but weaken



(a) first RGB-IR example (b) second RGB-IR example (c) first RGB-NIR example (d) second RGB-NIR example

Fig. 17. Qualitative results of fusion methods and our F2M on RGB-IR and RGB-NIR data with non-rigid local parallaxes. MG, EI, and VIF (introduced in Section IV-D2) evaluate the fusion performance (larger value: better performance).



(a) first PET-MRI example (b) second PET-MRI example (c) first CT-MRI example (d) second CT-MRI example

Fig. 18. Qualitative comparisons of different fusion methods and our F2M on PET-MRI and CT-MRI image pairs.

TABLE IV
QUANTITATIVE COMPARISON OF THE FINE REGISTRATION AND FUSION MODULE (F2M) AND STATE-OF-THE-ART FUSION COMPETITORS

Tasks	Metrics	DenseFuse [54]	DIF-Net [55]	IFCNN [57]	MDLatLRR [56]	RFN-Nest [33]	U2Fusion [6]	MURF (F2M)
RGB-IR	MG	3.751±0.809	2.961±0.644	5.394±1.203	4.020±0.871	3.399±0.841	5.075±1.156	**5.872±1.283**
	EI	31.067±6.811	28.471±6.293	47.904±10.975	35.922±8.137	33.553±8.613	45.213±10.916	**50.595±11.298**
	VIF	0.244±0.032	0.228±0.033	**0.323±0.047**	0.289±0.041	0.252±0.026	0.313±0.049	0.301± 0.044
RGB-NIR	MG	4.271±1.734	3.938±1.587	6.382±2.607	4.353±1.690	3.092±1.129	5.211±2.241	**7.442±3.035**
	EI	32.526±12.190	34.536±12.844	49.228±18.557	35.912±12.808	30.723±10.620	47.340±17.730	**57.639±20.693**
	VIF	0.306±0.050	0.356±0.054	**0.399±0.051**	0.390±0.047	0.343±0.039	0.397±0.065	0.396±0.060
PET-MRI	MG	7.724±0.865	8.704±0.842	12.801±1.281	9.488±1.021	5.948±0.494	10.358±0.915	**21.198±2.334**
	EI	55.530±4.291	73.792±5.864	104.991±8.312	78.373±6.305	56.753±4.820	88.986±7.107	**141.983±11.852**
	VIF	0.302±0.018	**0.507±0.056**	0.339±0.013	0.444±0.036	0.356±0.034	0.346±0.020	0.347±0.025
CT-MRI	MG	5.453±1.312	5.323±1.071	9.674±2.277	5.845±1.344	3.886±0.620	7.326±1.337	**13.132±3.067**
	EI	40.767±9.060	45.042±8.478	80.608±18.639	50.781±11.508	38.161±5.969	63.907±11.698	**92.927±21.744**
	VIF	0.147±0.023	**0.224±0.033**	0.167±0.034	0.171±0.035	0.216±0.036	0.164±0.034	0.204±0.032

the textures of MRI images. Our results preserve both functional and structural information and do not suffer from color distortion such as DenseFuse and IFCNN. Finally, F2M not only preserves but also enhances textures, as shown in Fig. 18(a) and (c).

2) Quantitative Results: Quantitative comparisons are performed on 50 RGB-IR, 50 RGB-NIR, 20 PET-MRI, and 20 CT-MRI image pairs. Three metrics, including mean gradient (MG) [58], edge intensity (EI) [59], and visual information fidelity (VIF) [60] are used for quantitative evaluation. MG evaluates the mean gradient, reflecting texture details of the fused image. EI measures the gradient amplitude of the edge point. A larger EI image represents a higher image quality and more clearness [61]. VIF is consistent with human visual system by measuring the information fidelity of the fused image. First, source images and fused image are filtered and divided into several blocks, respectively. Then, the visual information with and without distortion is evaluated and VIF for each subband is calculated. Finally, the overall VIF is calculated.

As reported in Table IV, F2M achieves the optimal performances on MG and EI on all the tasks. It indicates that our results

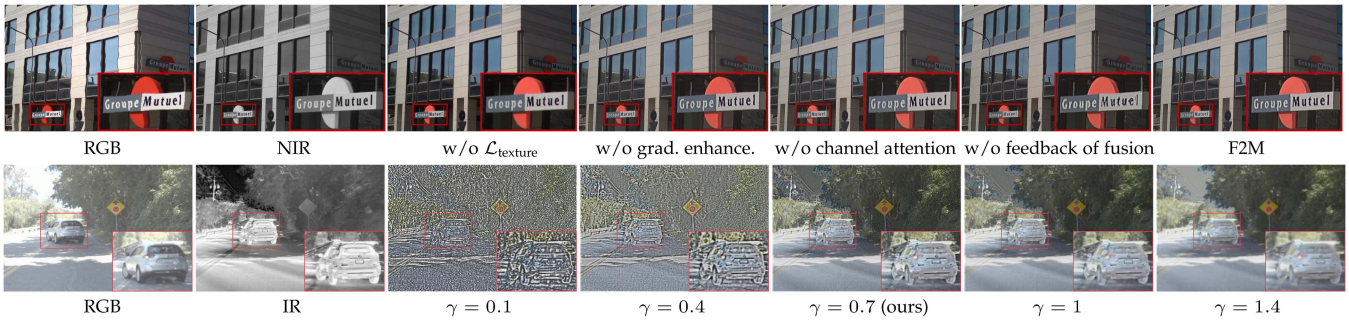


Fig. 19. The first row provides qualitative results of F2M and F2M without the texture loss ($\delta = 0$), gradient enhancement function ($\gamma = 1$), gradient channel attention, and feedback of the fused image ($\eta = 0$). The second row shows the qualitative results when setting γ in (11) to different values.

contain the most abundant texture details, highest image quality, and clearest contents. It is consistent with the characteristics demonstrated by the qualitative results. For VIF, F2M is inferior to some of the competitors. The reasons lie in two aspects. First, there are some local non-rigid offsets in the RGB-IR and RGB-NIR images. F2M corrects the offsets through the deformation block while the competitors do not. The correction naturally leads to a reduction in the similarity between the fused image and the source image before the deformation. It is reflected as a reduction in VIF. Second, our fusion method not only aims to retain the information in source images into the fused image but also enhance the textures to provide higher image quality. The enhancement operation inevitably leads to the differences between the information presented in the fusion result and the original information. These two factors lead our F2M to perform worse than some of the competitors on VIF, which measures the similarity between the fused image and source images.

3) Ablation Study and Hyper-Parameter Analysis: In F2M, to improve fusion performance, we design the texture loss $\mathcal{L}_{\text{texture}}$ and the gradient enhancement function in (11), and the gradient channel attention mechanism. Moreover, to improve registration accuracy, we rely on the characteristic of the fused image for feedback. This section validates the effectiveness of these settings. Also as the preliminary version (RFNet [7]) is designed for RGB-NIR data, we take RGB-NIR images to perform the ablation study. Then, the results can also be compared with RFNet.

As shown in Fig. 19, when the texture loss is not applied, the network fuses source images without emphasis and considering information quality. The fused image shows blur texture details. By using the texture loss to preserves sharper source textures, the result exhibits higher image quality, as illustrated in the fused image w/o gradient enhancement, w/o channel attention, and F2M. Comprehensively, the result of F2M contains more textures than other fusion results. It demonstrates that the gradient enhancement function and the channel attention introduce more scene details into the fused image. For the registration setting, we ablate the feedback of fusion by setting η in (12) to 0. Then, F2M only relies on the inverse deformation field for optimization. The results in Fig. 19 show that the single application of inverse deformation field results in an unsmooth deformation field. The constraint based on the gradient sparsity of the fused image further improves the accuracy.

The quantitative results tested on the same dataset with the same metrics in Section IV-D2 are reported in Table V. F2M shows the most details, clearest image quality, and highest information fidelity by achieving the best performances on MG, EI, and VIF. By ablating fusion-related settings, we find that the texture loss significantly improves the quality of fused image. The gradient enhancement function is of secondary importance. Lastly, the fusion performance is improved least with the gradient channel attention to revise the network. The results show that the combination of these settings can achieve the optimal performance, validating their effectiveness. With the feedback of fusion, these metrics are further improved with more accurate alignment.

Furthermore, we analysis the effect of different settings of γ in (11) on the fusion performance. As shown in Fig. 19, $\gamma = 0.1$ and $\gamma = 0.4$ result in over-sharpened textures that distort the visual effect of fused image. Contrarily, $\gamma > 1$ (e.g., $\gamma = 1.4$) blurs the original textures. By comparison, $\gamma = 0.7$ is more suitable for enhancement.

4) Exchanging Reference Images: In previous experiments, we set images in modality $\mathcal{Y}$ as reference images. Contrarily, we set images in modality $\mathcal{X}$ as reference images and apply deformations to images in modality $\mathcal{Y}$. Taking RGB-NIR images as an example, the results shown in Fig. 22 validate the generalization of F2M.

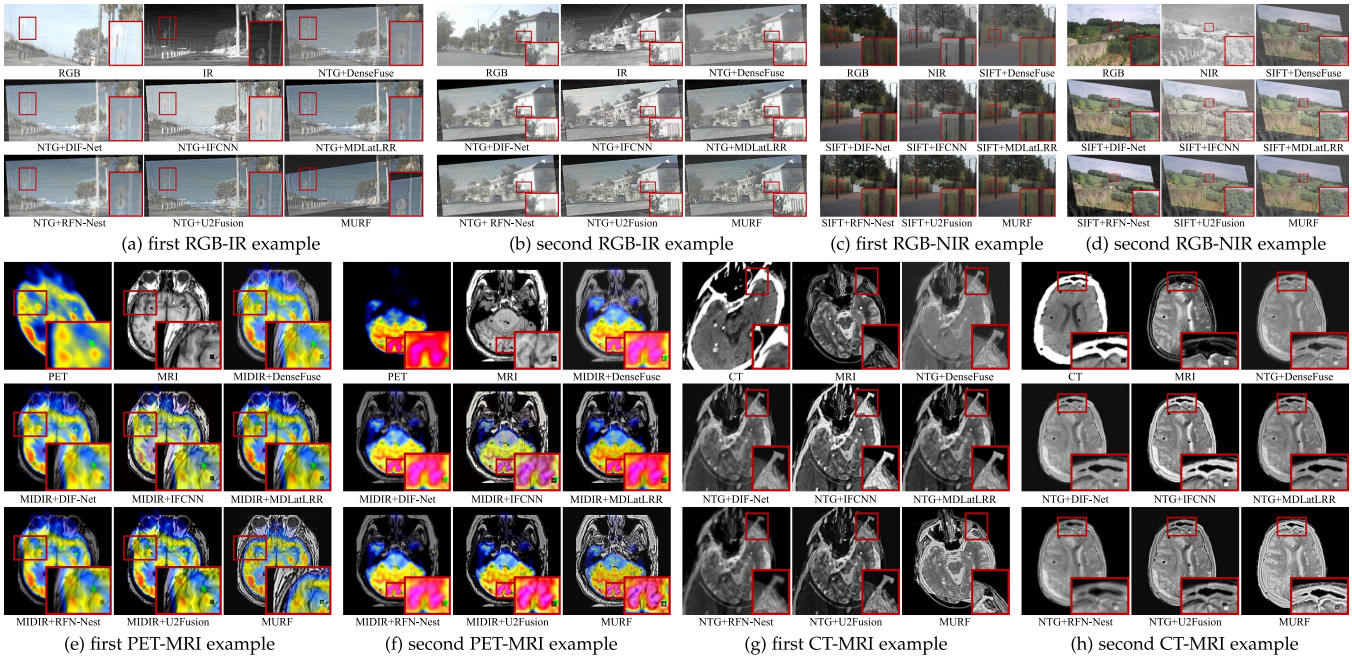
E. Combination of Registration and Fusion

We use the combination of different registration and fusion methods as competitors to evaluate the overall performance of MURF. As the comparison fusion methods cannot correct parallaxes, the registration competitors in Section IV-C which achieve the optimal performances in corresponding tasks are utilized for pre-processing. The registration method for RGB-IR, RGB-NIR, PET-MRI, and CT-MRI image pairs are NTG [49], SIFT [50], MIDIR [53], and NTG, respectively.

We analyze the qualitative results in Fig. 20 from two aspects. First, in Fig. 20(a), (c), (e), and (g), the comparison registration methods fail to completely eliminate the parallaxes. Some misalignments caused by deficient registration accuracy

TABLE V
QUANTITATIVE RESULTS OF F2M AND F2M WITHOUT THE TEXTURE LOSS ($\delta = 0$), GRADIENT ENHANCEMENT ($\gamma = 1$), GRADIENT CHANNEL ATTENTION, AND FEEDBACK OF FUSION. THE COMPARISON WITH F2M IN THE PRELIMINARY VERSION (RFNET) IS ALSO PROVIDED

	w/o $\mathcal{L}_{\text{texture}}$	w/o grad. enhance.	w/o grad. channel attention	w/o feedback of fusion	F2M of RFNet	F2M of MURF
MG	3.852±1.707	6.334±2.788	6.573±2.847	7.411±3.065	6.986±2.898	**7.442±3.035**
EI	35.728±13.815	50.294±18.482	51.880±19.010	56.538±20.226	55.285±20.561	**57.639±20.693**
VIF	0.351±0.061	0.375±0.060	0.385±0.060	0.382±0.056	0.243±0.071	**0.396±0.060**



(a) first RGB-IR example (b) second RGB-IR example (c) first RGB-NIR example (d) second RGB-NIR example

(e) first PET-MRI example (f) second PET-MRI example (g) first CT-MRI example (h) second CT-MRI example

Fig. 20. Overall performance comparison of the proposed method and the combination of state-of-the-art registration methods and fusion methods on unaligned RGB-IR, RGB-NIR, PET-MRI, and CT-MRI image pairs.

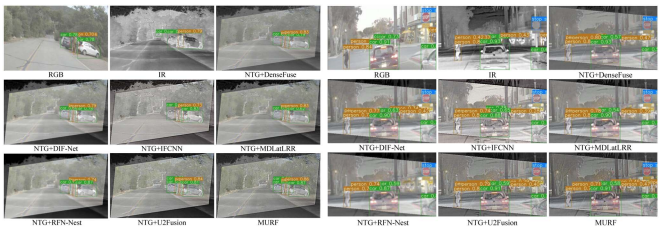


Fig. 21. Detection results on RGB-IR images, and different registration and fusion results.

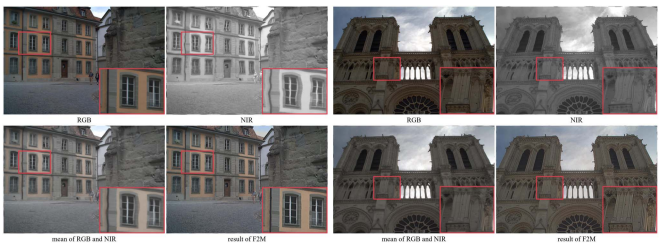


Fig. 22. Qualitative results when we contrarily set images of modality X (*e.g.*, RGB) as reference and correct the deformations in images of modality Y (*e.g.*, NIR).

remain in the results, causing incorrect position correspondence, overlapping shadows, and blurred results. By comparison, our coarse-to-fine fashion and multi-scale strategy correct the misalignments and represent correct correspondence and clearer textures. Second, our fusion results exhibit the most abundant textures. In Fig. 20(b), (d), (f), and (h), the blurred textures in one source image affect the clarity of competitors, particularly evident in PET-MRI and CT-MRI images. Our results preserve and enhance the sharper textures, which provide a more detailed portrayal of scenes and are suitable for human visual perception system.

External Verification of Object Detection. To evaluate the practical benefits of image fusion and its improved performance, external verification of object detection is performed. As there are no significant detection targets in medical images and RGB-NIR images are not synchronized (not taken at the same time), the verification is performed on RGB-IR images with YOLOv7 [62] as the detector. As shown in Fig. 21, the detector fails to detect all the thermal targets in source images, such as the three cars in the first example and the persons in the second example. However, after fusion, the three cars and more persons are successfully detected in some fused images. Besides, the confidence levels of some targets are increased in

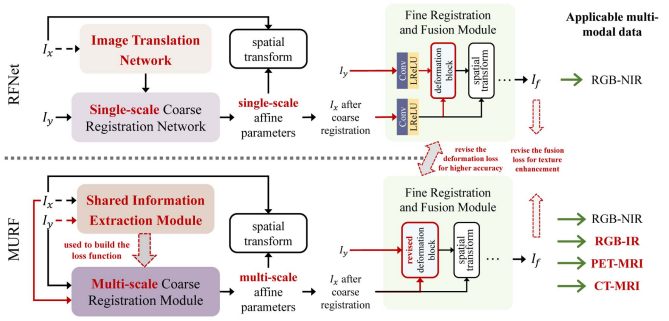


Fig. 23. Diagram showing the differences between the previous version (RFNet) and the proposed MURF in highlight. Different or improved parts are highlighted in red.

some results. These phenomenon verify the practical benefit of image fusion.

Comparing the detection results, MURF shows two advantages. First, more targets are detected in our results, as evidenced by the three cars in the first example and especially multiple persons below the stop sign in the second example. These targets are not detected in some competitors. Second, with the improvement of fusion performance, some confidence levels are increased in our results. In the first example, the confidence levels of the car and person in our result are higher than the source images and competitors.

F. MURF versus *RFNet*

The preliminary of MURF is RFNet [7]. The improvements are stated in Section I. MURF is applicable to coarse registration of not only street scenes as RFNet but also nature scenes. The differences and improvements are summarized as Fig. 23. The technical improvements are in four aspects.

The first two improvements are for coarse registration: *i)* RFNet employs image translation to eliminate modal variances. The registration loss is defined on the translated image and the reference one. In MURF, the registration loss is based on the extracted shared information illustrated in Fig. 11; *ii)* We revise the single-scale registration to multi-scale progressive registration. To validate the effectiveness, the qualitative comparison is performed on a street scene and a nature scene. As shown in the street scene in Fig. 24(a), RFNet and MURF can both correct the offsets as their results show minor deviations. However, for the nature scenes in Fig. 24(b), RFNet shows a distinct disadvantage, manifested in lower registration accuracy. The reason is that different types of scenes show distinct inter-modal structure differences. Street scenes contain significant structures which do not vary between modalities. Nature scenes may contain obvious inter-modal structure differences (e.g., grassland and forest) or lack salient structures (e.g., water). In image translation, the target of reducing inter-modal structure differences is basically achieved by re-styling existing contents. It is almost impossible to generate contents that do not originally exist. In MURF, we extract the shared information rather than generating structures that do not originally exist. Therefore, MURF shows higher registration accuracy and broader application scenarios. The

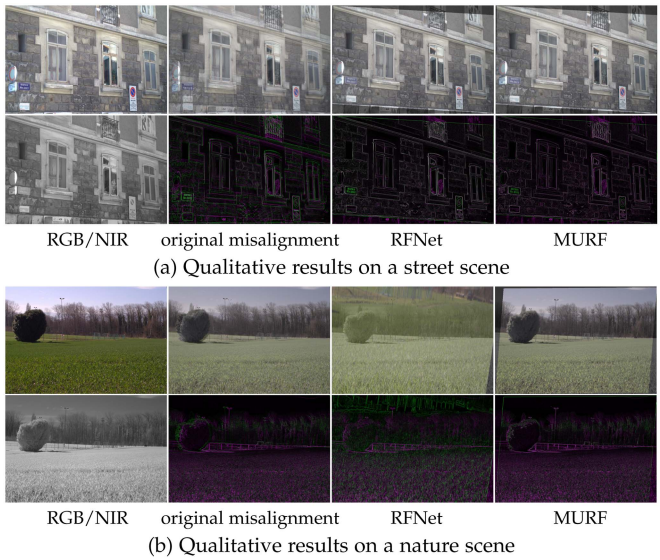


Fig. 24. Qualitative coarse registration comparison of RFNet and MURF. Second column: superimposed source images (second row: superimposed gradients); other columns: superimposed deformed RGB and NIR images.

TABLE VI
REGISTRATION COMPARISON OF RFNET AND MURF

Metrics	RMSE	MAE	MEE
RFNet	23.981±29.483	52.183±62.000	25.809±33.777
MURF	8.255±16.133	18.378±35.641	7.178±10.470

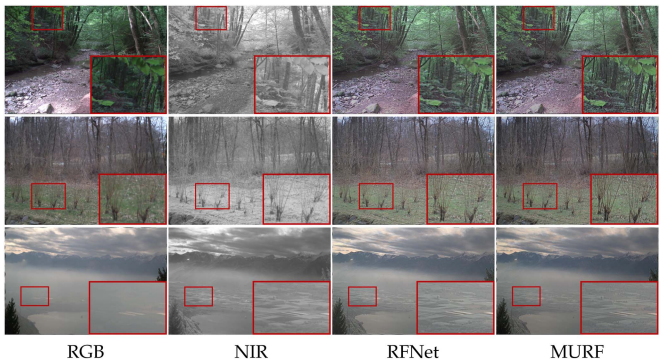


Fig. 25. Qualitative comparison of F2M in RFNet and MURF.

quantitative comparison of 75 unaligned RGB-NIR image pairs reported in Table VI also shows that MURF achieves better registration performance.

The other two technical improvements are in F2M: *i)* RFNet merely relies on the feedback of fusion to correct offsets while it may generate incorrect deformation fields. MURF not only relies on the feedback but also uses the inverse deformation field to improve the accuracy; *ii)* MURF not only preserves the source textures as in RFNet but also enhances the original textures with poor visibility. To validate these improvements, the qualitative results on typical RGB-NIR image pairs are shown in Fig. 25. It demonstrates two advantages of MURF over RFNet. In the first two examples, MURF achieves higher fine-registration accuracy

TABLE VII
PARAMETER COMPARISON OF DEEP LEARNING-BASED METHODS

Methods	Space transform		Registration			Fusion								
	RFNet (TransNet)	MURF (SIEM)	MIDIR	RFNet (AffineNet)	MURF (AffineNet)	DenseFuse	DIF-Net	IFCNN	RFN-Nest	U2Fusion	RFNet (F2M)		MURF (F2M)	
											Deformation	Fusion	Deformation	Fusion
Param	2237569	528626	83778	1428614	960998	74195	22465	83587	7524249	659217	579714	19841	243746	26269

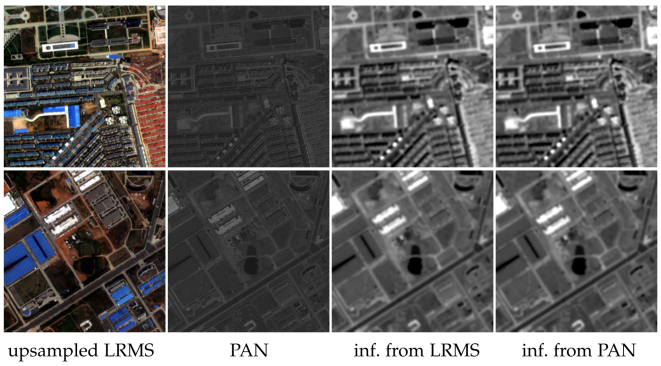


Fig. 26. Qualitative results of shared information extracted from the upsampled LRMS and PAN images, respectively.

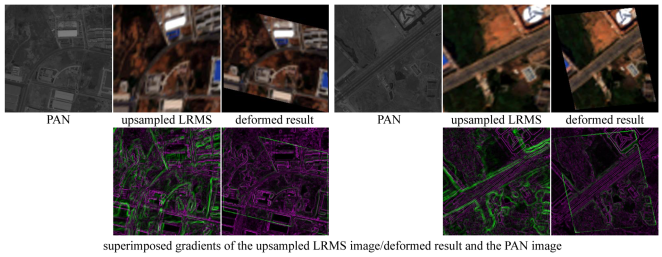


Fig. 27. Qualitative results of the registration of the upsampled LRMS and PAN images through MCRM in MURF.

than RFNet. In the last example, some regions are blurred and fuzzy in RFNet while they are sharper and more specific in MURF. Comparing Tables IV and V, RFNet contains more abundant texture details, higher image quality, and clearer scene contents than the comparison fusion methods. On this basis, MURF further improves the mean gradient, edge intensity, and visual information fidelity of the fused images. It further proves the superiority of our fusion performance.

G. Complexity Comparison

We perform a complexity comparison of all the comparison methods, our preliminary version (RFNet), and MURF in terms of the number of parameters. As reported in Table VII, when applying space transform to reduce modal variances, SIEM in MURF uses fewer parameters than TransNet in RFNet. For image registration, MURF uses more parameters than the comparison method, but achieves higher registration accuracy. Besides, it also shows fewer parameters than the registration network in RFNet. When purely comparing fusion-related parameters, MURF shows obvious advantages. Its complexity is only inferior to DIF-Net and RFNet. Most parameters in F2M exist in the deformation block, which is also advantageous over that in RFNet. In general, MURF has less computational complexity than RFNet.

V. FUTURE IMPROVEMENTS

Extension to Pansharpening. We investigate the possibility and effectiveness of the proposed method for pansharpening, i.e., fusing a low-resolution multispectral (LRMS) image and a high-resolution panchromatic (PAN) image. We evaluate on QuickBird dataset. First, the results of extracting shared information through SIEM are shown in Fig. 26. It indicates that our SIEM can effectively capture the shared information, which is more consistent than source images. Second, the registration results shown in Fig. 27 indicate that our MCRM can correct global rigid deformations in remote sensing images. Third, considering fusion, MS images contain more channels and not applicable for conversion to YCbCr space. Pansharpening requires unique and strict retention of spectral information. It is significantly different from previous tasks. Thus, F2M cannot yet be applied to pansharpening and it remains a future improvement.

Dealing With More Input Modalities. We consider the condition when processing $N(N > 2)$ input images of different modalities through this method. For registration, a reference image should be selected and the other $N - 1$ source images should be individually aligned with this image. For fusion, as F2M contains the gradient channel attention block, it can only fuse two images at a time. Thus, other input images should be fused with the fused image one by one. Therefore, the total computational cost of processing N input images is $N - 1$ times that of processing two images. A future improvement could be to reduce the computational complexity when dealing with more inputs.

VI. CONCLUSION

In this article, we propose a novel approach realizing multimodal image registration and fusion in a mutually reinforcing framework, termed as MURF. It breaks through the bottleneck that existing fusion methods are only applicable to aligned source images. MURF comprises three modules: shared information extraction module (SIEM), multi-scale coarse registration module (MCRM), and fine registration and fusion module (F2M). The image registration is handled in a coarse-to-fine approach. For coarse registration, SIEM first transforms multimodal images into mono-modal information to eliminate modal variances. Based on it, MCRM progressively corrects global rigid parallaxes through multi-scale affine transformation. The fine registration and fusion are realized in a single module, which further improves registration accuracy and fusion performance. For image fusion, we attempt to integrate texture enhancement

in addition to source information retention. The registration and fusion experiments on four multi-modal tasks validate the effectiveness and universality of the proposed method.

REFERENCES

[1] L. Tang, J. Yuan, and J. Ma, "Image fusion in the loop of high-level vision tasks: A semantic-aware real-time infrared and visible image fusion network," *Inf. Fusion*, vol. 82, pp. 28–42, 2022.

[2] N. LaHaye, M. J. Garay, B. D. Bue, H. El-Askary, and E. Linstead, "A quantitative validation of multi-modal image fusion and segmentation for object detection and tracking," *Remote Sens.*, vol. 13, no. 12, 2021, Art. no. 2364.

[3] Z. Huang, C. Lv, Y. Xing, and J. Wu, "Multi-modal sensor fusion-based deep neural network for end-to-end autonomous driving with scene understanding," *IEEE Sensors J.*, vol. 21, no. 10, pp. 11 781–11 790, May 2021.

[4] Z. Liu, E. Blasch, Z. Xue, J. Zhao, R. Laganiere, and W. Wu, "Objective assessment of multiresolution image fusion algorithms for context enhancement in night vision: A comparative study," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 1, pp. 94–109, Jan. 2012.

[5] H. Zhang, H. Xu, X. Tian, J. Jiang, and J. Ma, "Image fusion meets deep learning: A survey and perspective," *Inf. Fusion*, vol. 76, pp. 323–336, 2021.

[6] H. Xu, J. Ma, J. Jiang, X. Guo, and H. Ling, "U2Fusion: A unified unsupervised image fusion network," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 1, pp. 502–518, Jan. 2022.

[7] H. Xu, J. Ma, J. Yuan, Z. Le, and W. Liu, "RFNet: Unsupervised network for mutually reinforcing multi-modal image registration and fusion," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 19 679–19 688.

[8] 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, no. 1, pp. 23–79, 2021.

[9] 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.

[10] J. Chen et al., "WLD: A robust local image descriptor," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1705–1720, Sep. 2010.

[11] C. Wachinger and N. Navab, "Structural image representation for image registration," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2010, pp. 23–30.

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

[13] F. Zhao, Q. Huang, and W. Gao, "Image matching by normalized cross-correlation," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2006, pp. 729–732.

[14] B. B. Avants, C. L. Epstein, M. Grossman, and J. C. Gee, "Symmetric diffeomorphic image registration with cross-correlation: Evaluating automated labeling of elderly and neurodegenerative brain," *Med. Image Anal.*, vol. 12, no. 1, pp. 26–41, 2008.

[15] J. Luo and E. E. Konofagou, "A fast normalized cross-correlation calculation method for motion estimation," *IEEE Trans. Ultrason., Ferroelectr, Freq. Control*, vol. 57, no. 6, pp. 1347–1357, Jun. 2010.

[16] C. Studholme, D. L. Hill, and D. J. Hawkes, "An overlap invariant entropy measure of 3D medical image alignment," *Pattern Recognit.*, vol. 32, no. 1, pp. 71–86, 1999.

[17] C. Studholme, C. Drapaca, B. Iordanova, and V. Cardenas, "Deformation-based mapping of volume change from serial brain MRI in the presence of local tissue contrast change," *IEEE Trans. Med. Imag.*, vol. 25, no. 5, pp. 626–639, May 2006.

[18] D. Loeckx, P. Slagmolen, F. Maes, D. Vandermeulen, and P. Suetens, "Nonrigid image registration using conditional mutual information," *IEEE Trans. Med. Imag.*, vol. 29, no. 1, pp. 19–29, Jan. 2010.

[19] C. Wang, G. Papanastasiou, A. Chartsias, G. Jacenkow, S. A. Tsaftaris, and H. Zhang, "FIRE: Unsupervised bi-directional inter-modality registration using deep networks," 2019, *arXiv:1907.05062*.

[20] M. Arar, Y. Ginger, D. Danon, A. H. Bermano, and D. Cohen-Or, "Unsupervised multi-modal image registration via geometry preserving image-to-image translation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 13 410–13 419.

[21] K. He, D. Zhou, X. Zhang, and R. Nie, "Multi-focus: Focused region finding and multi-scale transform for image fusion," *Neurocomputing*, vol. 320, pp. 157–170, 2018.

[22] G. Li, Y. Lin, and X. Qu, "An infrared and visible image fusion method based on multi-scale transformation and norm optimization," *Inf. Fusion*, vol. 71, pp. 109–129, 2021.

[23] Q. Wei, J. Bioucas-Dias, N. Dobigeon, and J.-Y. Tourneret, "Hyperspectral and multispectral image fusion based on a sparse representation," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 7, pp. 3658–3668, Jul. 2015.

[24] Y. Liu, S. Liu, and Z. Wang, "A general framework for image fusion based on multi-scale transform and sparse representation," *Inf. Fusion*, vol. 24, pp. 147–164, 2015.

[25] H. Xu, M. Qin, S. Chen, Y. Zheng, and J. Zheng, "Hyperspectral-multispectral image fusion via tensor ring and subspace decompositions," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 8823–8837, 2021.

[26] F. Meng, B. Guo, M. Song, and X. Zhang, "Image fusion with saliency map and interest points," *Neurocomputing*, vol. 177, pp. 1–8, 2016.

[27] J. Ma, Z. Zhou, B. Wang, and H. Zong, "Infrared and visible image fusion based on visual saliency map and weighted least square optimization," *Infrared Phys. Technol.*, vol. 82, pp. 8–17, 2017.

[28] S. P. Yadav and S. Yadav, "Image fusion using hybrid methods in multimodality medical images," *Med. Biol. Eng. Comput.*, vol. 58, no. 4, pp. 669–687, 2020.

[29] N. Alseelawi, H. T. Hazim, and H. T. Salim ALRikabi, "A novel method of multimodal medical image fusion based on hybrid approach of NSCT and DTCWT," *Int. J. Online Biomed. Eng.*, vol. 18, no. 3, pp. 114–133, 2022.

[30] J. Ma, C. Chen, C. Li, and J. Huang, "Infrared and visible image fusion via gradient transfer and total variation minimization," *Inf. Fusion*, vol. 31, pp. 100–109, 2016.

[31] Q. Du, H. Xu, Y. Ma, J. Huang, and F. Fan, "Fusing infrared and visible images of different resolutions via total variation model," *Sensors*, vol. 18, no. 11, 2018, Art. no. 3827.

[32] H. Li, X.-J. Wu, and T. Durrani, "NestFuse: An infrared and visible image fusion architecture based on nest connection and spatial/channel attention models," *IEEE Trans. Instrum. Meas.*, vol. 69, no. 12, pp. 9645–9656, Dec. 2020.

[33] H. Li, X.-J. Wu, and J. Kittler, "RFN-Nest: An end-to-end residual fusion network for infrared and visible images," *Inf. Fusion*, vol. 73, pp. 72–86, 2021.

[34] H. Xu, X. Wang, and J. Ma, "DRF: Disentangled representation for visible and infrared image fusion," *IEEE Trans. Instrum. Meas.*, vol. 70, 2021, Art. no. 5006713.

[35] H. Xu and J. Ma, "EMFusion: An unsupervised enhanced medical image fusion network," *Inf. Fusion*, vol. 76, pp. 177–186, 2021.

[36] L. Tang, J. Yuan, H. Zhang, X. Jiang, and J. Ma, "PIAFusion: A progressive infrared and visible image fusion network based on illumination aware," *Inf. Fusion*, vol. 83, pp. 79–92, 2022.

[37] Y. Long, H. Jia, Y. Zhong, Y. Jiang, and Y. Jia, "RXDNFuse: A aggregated residual dense network for infrared and visible image fusion," *Inf. Fusion*, vol. 69, pp. 128–141, 2021.

[38] J. Ma, W. Yu, P. Liang, C. Li, and J. Jiang, "FusionGAN: A generative adversarial network for infrared and visible image fusion," *Inf. Fusion*, vol. 48, pp. 11–26, 2019.

[39] J. Ma et al., "Infrared and visible image fusion via detail preserving adversarial learning," *Inf. Fusion*, vol. 54, pp. 85–98, 2020.

[40] J. Ma, H. Xu, J. Jiang, X. Mei, and X.-P. Zhang, "DDcGAN: A dual-discriminator conditional generative adversarial network for multi-resolution image fusion," *IEEE Trans. Image Process.*, vol. 29, pp. 4980–4995, 2020.

[41] J. Ma, H. Zhang, Z. Shao, P. Liang, and H. Xu, "GANMcC: A generative adversarial network with multiclassification constraints for infrared and visible image fusion," *IEEE Trans. Instrum. Meas.*, vol. 70, 2021, Art. no. 5005014.

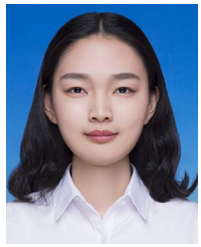
[42] M. C. El-Mezouar, K. Kpalma, N. Taleb, and J. Ronsin, "A pan-sharpening based on the non-subsampled contourlet transform: Application to worldview-2 imagery," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 7, no. 5, pp. 1806–1815, May 2014.

[43] H. Li, L. Jing, Y. Tang, and H. Ding, "An improved pansharpening method for misaligned panchromatic and multispectral data," *Sensors*, vol. 18, no. 2, 2018, Art. no. 557.

[44] A. V. D. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," 2018, *arXiv:1807.03748*.

[45] N. Pielawski et al., "CoMIR: Contrastive multimodal image representation for registration," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2020, pp. 18 433–18 444.

[46] X. Huang and S. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 1501–1510.
[47] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.
[48] L. Xu, S. Zheng, and J. Jia, "Unnatural L0 sparse representation for natural image deblurring," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 1107–1114.
[49] S.-J. Chen, H.-L. Shen, C. Li, and J. H. Xin, "Normalized total gradient: A new measure for multispectral image registration," *IEEE Trans. Image Process.*, vol. 27, no. 3, pp. 1297–1310, Mar. 2018.
[50] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.
[51] S. Kim, D. Min, B. Ham, S. Ryu, M. N. Do, and K. Sohn, "DASC: Dense adaptive self-correlation descriptor for multi-modal and multi-spectral correspondence," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 2103–2112.
[52] S. Kim, D. Min, B. Ham, M. N. Do, and K. Sohn, "DASC: Robust dense descriptor for multi-modal and multi-spectral correspondence estimation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 9, pp. 1712–1729, Sep. 2017.
[53] H. Qiu, C. Qin, A. Schuh, K. Hammernik, and D. Rueckert, "Learning diffeomorphic and modality-invariant registration using B-splines," in *Proc. 4th Conf. Med. Imag. Deep Learn.*, 2021, pp. 645–664.
[54] H. Li and X.-J. Wu, "DenseFuse: A fusion approach to infrared and visible images," *IEEE Trans. Image Process.*, vol. 28, no. 5, pp. 2614–2623, May 2019.
[55] H. Jung, Y. Kim, H. Jang, N. Ha, and K. Sohn, "Unsupervised deep image fusion with structure tensor representations," *IEEE Trans. Image Process.*, vol. 29, pp. 3845–3858, 2020.
[56] H. Li, X.-J. Wu, and J. Kittler, "MDLatLRR: A novel decomposition method for infrared and visible image fusion," *IEEE Trans. Image Process.*, vol. 29, pp. 4733–4746, 2020.
[57] Y. Zhang, Y. Liu, P. Sun, H. Yan, X. Zhao, and L. Zhang, "IFCNN: A general image fusion framework based on convolutional neural network," *Inf. Fusion*, vol. 54, pp. 99–118, 2020.
[58] J. Ma, Y. Ma, and C. Li, "Infrared and visible image fusion methods and applications: A survey," *Inf. Fusion*, vol. 45, pp. 153–178, 2019.
[59] X. Luo, Z. Zhang, C. Zhang, and X. Wu, "Multi-focus image fusion using HOSVD and edge intensity," *J. Vis. Commun. Image Representation*, vol. 45, pp. 46–61, 2017.
[60] Y. Han, Y. Cai, Y. Cao, and X. Xu, "A new image fusion performance metric based on visual information fidelity," *Inf. Fusion*, vol. 14, no. 2, pp. 127–135, 2013.
[61] H. M. El-Hoseny, E.-S. M. El Rabaie, W. Abd Elrahman, and F. E. Abd El-Samie, "Medical image fusion techniques based on combined discrete transform domains," in *Proc. Nat. Radio Sci. Conf.*, 2017, pp. 471–480.
[62] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," 2022, *arXiv:2207.02696*.



Han Xu received the BS degree from the Electronic Information School, Wuhan University, Wuhan, China, in 2018. She is currently working toward the PhD degree with the Multi-spectral Vision Processing Lab, Electronic Information School, Wuhan University. She has first-authored more than 10 refereed journal and conference papers, including *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *IEEE Transactions on Image Processing*, *Information Fusion*, CVPR, AAAI, IJCAI, *etc*. Her current research interests include computer vision and image processing.

Jiteng Yuan received the BS degree from the Electronic Information School, Wuhan University, Wuhan, China, in 2021. He is currently working toward the master degree with the Multi-Spectral Vision Processing Lab, Electronic Information School, Wuhan University. His research interests include computer vision and image processing.

Jiayi Ma (Senior Member, IEEE) received the BS degree in information and computing science and the PhD degree in control science and engineering from the Huazhong University of Science and Technology, Wuhan, China, in 2008 and 2014, respectively. He is currently a professor with the Electronic Information School, Wuhan University. He has authored or co-authored more than 300 refereed journal and conference papers, including *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *IEEE Transactions on Image Processing*, *International Journal of Computer Vision*, CVPR, ICCV, ECCV, *etc*. His research interests include computer vision, machine learning, and pattern recognition. He has been identified in the 2019–2022 Highly Cited Researcher lists from the Web of Science Group. He is an area editor of the *Information Fusion*, and an associate editor of the *IEEE/CAA Journal of Automatica Sinica* and *Neurocomputing*.