RFNet: Unsupervised Network for Mutually Reinforcing Multi-modal Image Registration and Fusion

Han Xu[1], Jiayi Ma[1*], Jiteng Yuan[1], Zhuliang Le[1], and Wei Liu[2]
[1] Wuhan University, China [2] Tencent Data Platform, China
{xu_han,yuanjiteng,lezhuliang}@whu.edu.cn, jyma2010@gmail.com, wl2223@columbia.edu

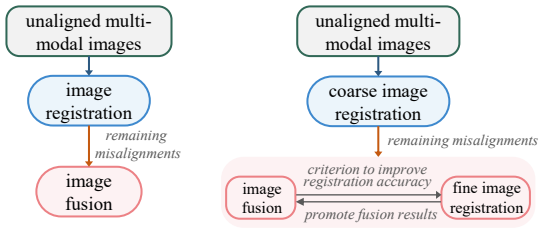
Abstract

In this paper, we propose a novel method to realize multi-modal image registration and fusion in a mutually reinforcing framework, termed as RFNet. We handle the registration in a coarse-to-fine fashion. For the first time, we exploit the feedback of image fusion to promote the registration accuracy rather than treating them as two separate issues. The fine-registered results also improve the fusion performance. Specifically, for image registration, we solve the bottlenecks of defining registration metrics applicable for multi-modal images and facilitating the network convergence. The metrics are defined based on image translation and image fusion respectively in the coarse and fine stages. The convergence is facilitated by the designed metrics and a deformable convolution-based network. For image fusion, we focus on texture preservation, which not only increases the information amount and quality of fusion results but also improves the feedback of fusion results. The proposed method is evaluated on multi-modal images with large global parallaxes, images with local misalignments and aligned images to validate the performances of registration and fusion. The results in these cases demonstrate the effectiveness of our method.

1. Introduction

Multi-modal image fusion aims to merge the information from different imaging modalities to generate a single image with rich information and high quality. Because the fused images can describe scenes comprehensively by merging the complementary information, image fusion serves as a powerful tool for wide applications, such as security, remote sensing, clinical treatment, *etc*.

As multi-modal images are taken from different devices/sensors, it inevitably leads to parallaxes due to biased positions, angles, *etc*. However, almost all the fusion methods fail to consider parallaxes. They require an accurate registration before fusion can take place, as shown in Fig. 1(a). Unfortunately, the diversity between different modalities poses a great challenge to improve the registration accuracy, still resulting in mitigated misalignments in the pre-registered images. *When registration and fusion are separate issues, existing fusion methods have to "tolerate" rather than "fight" the pre-registration misalignments.* Thus, multi-modal image registration and fusion become an urgent issue for the practical application of image fusion.



(a) Separate issues (b) Proposed method

Figure 1. Treating registration and fusion as separate issues in existing methods and the proposed mutually reinforcing framework.

Meanwhile, in existing separate branches, image fusion is a downstream task of registration, and fails to provide feedbacks to improve registration accuracy. 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. When the fused image is registered with either source image, the alleviated modal diversity reduces the difficulty of registration. Second, the misalignments in fused images undoubtedly lead to more but repeated salient structures, *i.e*., dense gradients. By comparison, an accurate registration encourages the sparseness of gradients. Thus, the gradient sparsity of fusion results can act as a criterion to improve registration accuracy in a feedback fashion without losing the scene information in source images. Third, the fused images retain the obvious salient structures in a single image and discard some superfluous and useless information during the fusion process. It reduces the negative impact of superfluous information on image registra-

*Corresponding author

tion. When image fusion helps to eliminate misalignments, the more precisely aligned data further promotes fusion results. As a result, these two tasks can be mutually reinforced in this way, as illustrated in Fig. 1(b).

Specifically to the individual solution of each task, either image registration or fusion has its own bottlenecks. For image registration, it is a difficult problem to *develop appropriate registration metrics or evaluation ways adaptive for multi-modal data*. The other important issue is to ensure that *the designed registration constraints should be practical for deep network optimization though gradient descent.* For image fusion, a general purpose is to enable fused images present the most amount of information, partly represented by gradients. Moreover, as stated above, the gradients of fused images play a critical role in eliminating misalignments. Combining these two aspects, fusion methods should *be dedicated to the retention of texture information*, which is consistent with both the fusion target and the feedback function of image fusion to image registration.

To address the limitations of prior works and unexplored issues, we explore multi-modal image registration and fusion in a mutually reinforcing framework. We propose an unsupervised network to realize it, termed as *RFNet*. The proposed framework is summarized as Fig. 1(b). The registration is handled in a coarse-to-fine approach. The coarse stage corrects the global parallaxes through an evaluation metric based on image translation. The coarse-registered results help generate meaningful but rough fused images. Image fusion and fine registration are integrated in a single network. Then, to correct local misalignments, we rely on the characteristics of fused images to optimize the deformation-related parts in this network. Finally, the network generates the fine-registered and fused image.

The main contributions of RFNet are summarized as follows: **i)** The problems of multi-modal image registration and fusion are mutually reinforced in our work. It is the first time that image fusion is exploited to promote multi-modal image registration accuracy through a deep neural network. **ii)** We focus on designing constraints to optimize the multi-modal registration performance. In the coarse stage, we apply image translation to build an image-level evaluation metric. An improved network architecture is proposed to help facilitate the network convergence. In the fine stage, the metric is designed based on the fusion results. **iii)** Considering the texture retention in image fusion, we adapt a gradient channel attention mechanism to adaptively adjust the channel-wise contributions of features. Besides, we design a gradient loss with bias. The network architecture and loss function are both based on the texture richness.

2. Related Works

Multi-modal Image Registration. Traditional registration methods include transformation- and measure-based ones. Transformation-based ones transfer images into a common space to exhibit better consistency [3, 10, 11, 26]. They manually analyse multi-modal characteristics and design constraints to impose consistency. Nevertheless, the optimization in these methods is thorny. Measure-based ones aim to measure the similarity with low sensitivity to modal variations. Representative methods utilize mutual information (MI), regional MI [23], *etc*., which are computationally intractable and not suitable for gradient descent [5]. Recently, deep learning-based methods have been proposed. For instance, Wang *et al*. [27] use a network to create modal-independent features while drawbacks of sparsity still exist. Closest to our work, Arar *et al*. learn a cross-modality translation [1]. However, the cooperative training of translation and registration networks increases the difficulty of optimizing registration network. In our work, we find that feeding translated images in the same domain into the network can improve the registration accuracy and speed up the convergence simultaneously. Besides, compared with existing registration networks [1,20], we employ deformable convolution in our network as it refers to the deformation in unregistered images for higher registration accuracy and stronger robustness. Most relevant to our work, SIRF [4] confirms that joint registration and fusion can definitely improve the results if they are combined properly. However, this work is realized in a tradition vectorial total variation model and designed for remote sensing images with restrictive local misalignments.

Multi-modal Image Fusion. Existing fusion methods are tailored to aligned images without regard to parallaxes. Focusing on fusion itself, traditional methods include six categories: methods based on multi-scale transform [16], sparse representation [28], subspace, saliency [21], hybrid methods, and others. They are devoted to designing decomposition ways and fusion strategies in a manual way while detailed and diverse designs make them more and more complex. To solve it, some deep learning-based methods are proposed [13, 15, 29]. Some of them do not pay attention to texture preservation and some generative adversarial network-based methods [7, 18, 19, 32] suffer from generating fake and blurred details. Even some methods concern the textures [17, 31], they preserve the textures according to the image modality rather than the actual textures of specific regions. In this work, we adapt a gradient-based attention mechanism and a gradient loss with bias to enhance texture retention. Moreover, the network blends the deformation which enables misalignment correction based on the preserved textures.

3. Proposed Method

We design an unsupervised network for mutually reinforcing multi-modal image registration and fusion, term as *RFNet*. The overall procedure is shown in Fig. 2, which

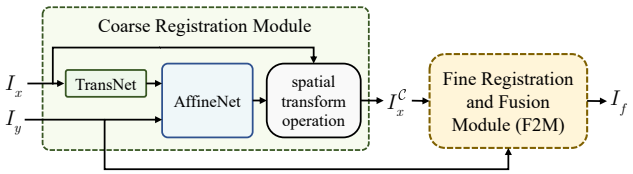


Figure 2. Overall pipeline of our RFNet. I_x and I_y are a paired but unaligned multi-modal images such as visible (VIS) and near-infrared (NIR) images. $I_x^{\mathcal{C}}$ is the coarse aligned I_x after coarse registration. I_f is the final aligned and fused image. TransNet is an image translation network to narrow modal differences. AffineNet is a network to generate the affine transformation parameters.

consists of two main parts. First, a coarse registration module performs the global correction based on the affine transformation model. Then, multi-modal images are roughly aligned except for some local parallaxes, where an affine model is not applicable. Second, the fine registration and fusion are realized in a unified module/network, termed as fine registration and fusion module (F2M).

3.1. Coarse Registration Module

Pipeline of the proposed coarse registration module is shown in Fig. 3. TransNet firstly transfers multi-modal images into the same domain (*i.e.*, translating I_x to $I_{x\to y}$). AffineNet takes $I_{x\to y}$ and I_y as input, and outputs the affine parameters to generate the deformation field for I_x.

3.1.1 Image Translation Network

TransNet is aimed at learning the image translation function $\mathcal{T}_x^y$, which denotes translating an image I_x in domain x to domain y by retaining content information. Thus, we use an encoder to embed I_x into the content space as $c_x = E_x(I_x)$ while removing the domain information. To ensure that c_x contains the content information, we map it back to domains through the decoders D_x and D_y, as shown in Fig. 4.

The result of mapping c_x back to domain x is expected to reconstruct I_x, *i.e.*, $I_x^{\text{recon}} = \mathcal{T}_x^x(I_x) = D_x(E_x(I_x))$. And the mapping result to domain y should be the translated I_x, *i.e.*, $I_{x\to y} = \mathcal{T}_x^y(I_x) = D_y(E_x(I_x))$. Similarly, for I_y in domain y, the reconstructed and translated results are $I_y^{\text{recon}} = \mathcal{T}_y^y(I_y)$ and $I_{y\to x} = \mathcal{T}_y^x(I_y)$.

To encourage encoders to extract content information and decoders to recover domain-related information, the reconstruction loss and the translation loss are defined as:

$$\begin{aligned}\mathcal{L}_{\text{recon}} &= \|I_x - I_x^{\text{recon}}\|_1 + \|I_y - I_y^{\text{recon}}\|_1,\\ \mathcal{L}_{\text{trans}} &= \|I_x - I_{y\to x}\|_1 + \|I_y - I_{x\to y}\|_1.\end{aligned} \tag{1}$$

The final loss function of TransNet is summarized as Eq. (2) with a hyper-parameter η controlling the trade-off:

$$\mathcal{L}_{\text{TransNet}} = \mathcal{L}_{\text{recon}} + \eta\mathcal{L}_{\text{trans}}. \tag{2}$$

Network Architecture. The network architecture of TransNet is shown in *Supplementary Material*. We use instance normalization rather than batch normalization as it performs a kind of style normalization [8]. To map different domains to a same content space, in addition to the designed loss functions, the weights of the last layers in encoders and the first layers in decoders are shared.

3.1.2 Affine Network

AffineNet learns to generate the corresponding affine transformation function $\mathcal{C}$. When feeding a pair of unaligned images $I_{x\to y}$ and I_y, it outputs the affine parameters $p_{\text{aff}} = \mathcal{C}(I_{x\to y}, I_y)$. According to p_{aff}, we generate a deformation field ϕ of size $H \times W \times 2$ by applying p_{aff} on a regular sampling grid. ϕ represents the deformation of all pixels in $I_{x\to y}$. Mathematically, the deformed $I_{x\to y}$ is denoted as:

$$I_{x\to y}^{\mathcal{C}}[i + \phi_{i,j,1}, j + \phi_{i,j,2}] = I_{x\to y}[i, j], \tag{3}$$

where i and j denote the position of pixels. Two channels of ϕ denote the deviation in vertical and horizontal directions, respectively. Considering that there may be some missing pixel values due to the different coordinate types, a re-sampler $\mathcal{S}$ is applied for the betterment of this step.

As described, the problem of multi-modal image registration has been transformed as the similarity between the deformed translated image $I_{x\to y}^{\mathcal{C}}$ and the source image I_y. Therefore, the loss function of AffineNet is defined to constrain their similarity. For ease of computational tractability and for weaker sensibility to linear changes in illumination amplitudes, we use normalized cross correlation (NCC) as similarity measure. The registration loss is thus defined as:

$$\mathcal{L}_{\text{coarse}} = -NCC(I_{x\to y}^{\mathcal{C}}, I_y), \tag{4}$$

where $NCC(s, g)$ is defined as:

$$NCC(s, g) = \frac{\mathbb{E}[(s - \mu_s) \odot (g - \mu_g)]}{\sqrt{\mathbb{E}[(s - \mu_s)^2]}\sqrt{\mathbb{E}[(g - \mu_g)^2]}}, \tag{5}$$

where $\mathbb{E}[x] = \frac{1}{HW}\sum_{i=1}^{H}\sum_{j=1}^{W} x_{i,j}$, with $x_{i,j}$ being the pixel of x in the i-th row and j-th column. μ_s and μ_g are the mean values of s and g. $\odot$ is the Hadamard product.

When the optimal deformation field ϕ is obtained, we perform the same spatial transform on I_x to generate the coarse aligned image $I_x^{\mathcal{C}}$ according to the way in Eq. (3).

Network Architecture. The network architecture of AffineNet is reported in *Supplementary Material*. For image registration, the region of corresponding objects may shift considerably in two unregistered images. Taking the long-distance parallax into account, large kernel sizes and deep network layers are necessary for wide receptive fields

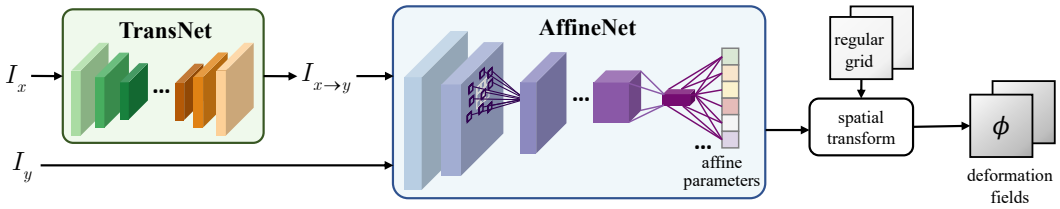


Figure 3. Procedure of the coarse registration module to generating the coarse deformation field.

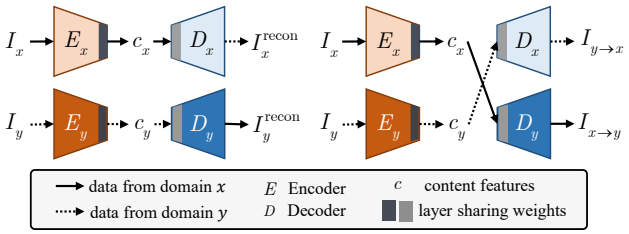


Figure 4. Multi-modal image translation configurations.

to alleviate this problem. Thus, deformable convolution layers are applied to replace the regular receptive fields in traditional convolution layers. Deformable convolution layers augment receptive fields with offsets, which are learned from additional convolution layers from preceding feature maps. Thus, it refers to deformations in unregistered images for higher registration accuracy and stronger robustness.

3.2. Mutually Reinforcing Fine Registration and Fusion Module (F2M)

In the first phase, F2M realizes the texture-focused image fusion, which is also the foundation of fine registration. The pipeline is shown in Fig. 5. We optimize the parameters in F2M for image fusion except those in the deformation block. The deformation block depends on initialized parameters to generate the deformation field, which automatically tends to be identical. In this case, I_f combines the scene information of $I_x^{\mathcal{C}}$ and I_y, and renders their parallaxes in a single image. Loss function is defined as:

$$\mathcal{L}_{\text{fus}} = \mathcal{L}_{\text{content}} + \delta \mathcal{L}_{\text{gradient}}, \tag{6}$$

where δ controls the trade-off between these two terms. $\mathcal{L}_{\text{content}}$ constrains the image-level similarity to merge the scene content, which is defined as:

$$\mathcal{L}_{\text{content}} = (1-\gamma)\|I_f - I_x^{\mathcal{C}}\|_1 + \gamma\|I_f - I_y\|_1. \tag{7}$$

As the NIR images usually contain more texture details than RGB images, γ is set to a value between 0.5 and 1.

As salient structures are usually presented in larger gradients, the gradient loss $\mathcal{L}_{\text{gradient}}$ is defined as:

$$\mathcal{L}_{\text{gradient}} = \left\| \nabla I_f - \frac{\nabla I_x^{\mathcal{C}} + \nabla I_y}{|\nabla I_x^{\mathcal{C}} + \nabla I_y|} \cdot \max(|\nabla I_x^{\mathcal{C}}|, |\nabla I_y|) \right\|_2, \tag{8}$$

where ∇ denotes the gradient of an image.

In the second phase, F2M realizes the fine registration based on the characteristics of fused images. In this phase, we fix the fusion-related parameters which have been optimized in the first phase and train the deformation block. The loss function considers the following three aspects. First, I_y is the fixed image providing the reference texture information. I_f retains the deformed gradients of $I_x^{\mathcal{C}}$. After the correct deformation, ∇I_f should demonstrate high consistency with ∇I_y. Thus, the first term constrains the consistency with the reference information. Second, it is easy to observe that any misalignments in I_f will decrease the sparsity of gradients. We use the second term to encourage the sparsity of ∇I_f and penalize the salient gradients that should be corrected. Third, it is clear that neighboring pixels should have similar deformations, intuitively represented by the smoothness of the deformation field. Otherwise, the scene structure will be distorted. We deploy a regularization term to prevent the deformation block from generating non-smooth deformation fields. Therefore, the loss function contains the following three terms:

$$\mathcal{L}_{\text{defor}} = \|\nabla I_f - \nabla I_y\|_1 + \|\nabla I_f\|_1 + \lambda \mathcal{L}_{\text{smooth}}, \tag{9}$$

where we use the l_1-norm as it encourages sparsity.

Specifically to $\mathcal{L}_{\text{smooth}}$, denoting the deformation as ϕ_f, the first order gradients of ϕ_f reflect the abrupt changes of the deformation. Besides, to avoid over-smoothing, inspired by [1], a bilateral filter [24] is used to assign variable weights to different first-order changes, defined as:

$$\mathcal{L}_{\text{smooth}} = \sum_{p_n \in \mathcal{R}} e^{-\alpha |I_f(p) - I_f(p_n)|} \cdot |\phi_f(p) - \phi_f(p_n)|, \tag{10}$$

where p is the position index of a pixel in I_f or ϕ_f. $\mathcal{R}$ denotes a set of neighbors of p. p_n represents the position index in this set. α is a coefficient and set to 0.5.

When the deformation block has been optimized, we once again perform the forward process of F2M entirely to generate the final aligned and fused image I_f.

Network Architecture. As shown in Fig. 5, we share the weights of the first three layers to ensure the intensity consistency of feature types from different modalities. It avoids the attenuation and diffusion of information in one source image compared to the other one. Otherwise, the attenuation and diffusion will cause the fake gradient sparsity and affect the improvement of registration performance.

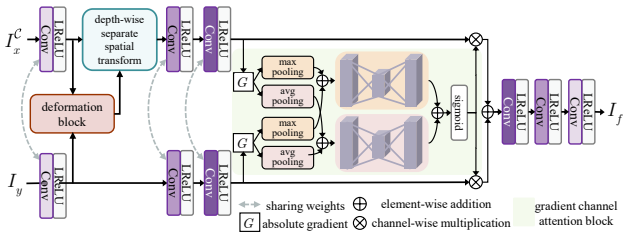


Figure 5. Pipeline of the fine registration and fusion module (F2M). "Conv": convolution layer with kernel size as 3×3 and stride as 1.

As the receptive fields grow with depth increases, a pixel in deeper feature maps corresponds to a larger region in the image, which are not conducive to improving the registration accuracy. Thus, we use the shallow features to explore and generate spatial deformation. The nonlinear mapping of the first convolution layer eliminates the pixel intensity differences between $I_x^{\mathcal{C}}$ and I_y. The deformation block generates the deformation fields (refer to *Supplementary Material* for details). Resampling, batch normalization [9], and residual blocks are used to apply to different deformations.

For texture preservation, we introduce the gradient channel attention block as in Fig. 5. 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 operations. Then, the two-branch results are added and fed into two individual multi-layer perceptrons to generate shared channel-wise attention weights. Then, several convolution layers map the features back to generate I_f.

4. Experiments

Implementation Details. The code of our method is implemented in TensorFlow. Experiments are conducted on an NVIDIA Geforce GTX Titan X GPU and 2.4 GHz Intel Core i5-1135 CPU. The parameters in all networks are updated with the Adam Optimizer [12]. The epoch of training coarse registration network is set to 100, and that of training F2Net is set to 30. The batch size is 4. The learning rate is set to 0.0004 with exponential decay. The hyper-parameters are set as: $\eta = 2$, $\delta = 100$, $\gamma = 0.7$, $\lambda = 0.1$. We build the training and test datasets based on a publicly available VIS-NIR Scene dataset [2][1]. Images are cropped into patches of size 384×384 and flipped for more training data.

4.1. Multi-modal Image Registration

We compare our coarse registration module with SOTA multi-modal registration methods, including traditional ones (*i.e.*, MI [25], DASC [10, 11], NTG [5], SCB [3]) and a deep learning-based method NeMAR [1]. For NeMAR, we retrain the model on our training dataset for 800 epoches.

[1] http://matthewalunbrown.com/nirscene/nirscene.html

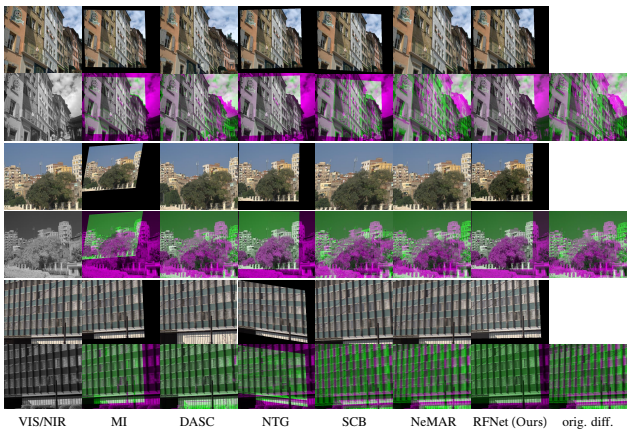


Figure 6. Registration results on three unaligned multi-modal image pairs. Under registration results, the deformed VIS image and NIR image are shown in pair to show their misalignments.

Qualitative Results. Qualitative results are shown in Fig. 6. In the first two groups, the proposed RFNet and NTG show more accurate registration results than others. MI and SCB perform almost exactly on the first pair while suffer a large registration error in the second pair. DASC shows severe geometric distortions, especially in the un-overlapped regions of two source images. NeMAR shows a slight improvement than unaligned images. In the third group, source images exhibit high structure similarity and repeatability in different regions. In this case, the proposed RFNet shows higher registration accuracy than comparative methods, including NTG. These results demonstrate that our method can outperform the SOTA methods.

Quantitative Evaluation. For quantitative evaluation, we build 5 pairs of point landmarks in each image pair (see *Supplementary Material* for illustration). The points in the deformed VIS image are expected to be in the same positions as those in the NIR image. Thus, we measure the Euclidean distance between the deformed source points and target points. We count the distances from three aspects, including root mean square error (RMSE), max square error (MAE) and median square error (MEE). Furthermore, we measure the image-level similarity between the deformed VIS and NIR images with peak signal-to-noise ratio (PSNR) and structural similarity index measure (SSIM). All the metrics are tested on 45 unaligned multi-modal image pairs and reported in Tab. 1. The coarse registration module of RFNet achieves the optimal results on RMSE, MAE and MEE. By comparison, MI and NTG achieve low means but high standard deviations as they perform well in some scenarios while not in others. DASC shows the optimal results in SSIM and PSNR because the results contain some incorrect information in un-overlapped regions. However,

Table 1. Quantitative comparisons of registration accuracy (mean and standard variation, red: best, blue: second best, cyan: third best).

Methods	unaligned VIS	MI [25]	DASC [10, 11]	NTG [5]	SCB [3]	NeMAR [1]	coarse registration module of RFNet
RMSE	47.223±12.729	15.756±53.284	17.214±10.115	9.942±23.807	37.384±25.209	43.747±12.426	3.972±4.064
MAE	74.652±17.786	25.347±79.450	40.752±22.688	18.221±42.164	67.760±39.807	69.616±17.498	5.209±2.079
MEE	66.199±18.287	22.102±76.505	15.989±14.240	13.430±35.279	49.605±36.250	61.059±17.910	3.475±3.336
SSIM	0.292±0.105	0.505±0.177	0.726±0.064	0.611±0.164	0.324±0.177	0.345±0.115	0.650±0.102
PSNR	11.768±1.996	14.837±3.672	17.582±3.052	15.349±3.453	12.784±3.429	12.179±2.139	15.043±1.073

in other results, the un-overlapped regions are black with little similarity with NIR images. In general, our method exhibits comparative registration performances.

4.2. Multi-modal Fusion and Our Fine Registration

This section focuses on evaluating the fusion and fine registration performances of our F2M. As the state-of-the-art (SOTA) fusion methods cannot deal with unaligned data, the registration method NTG [5] is utilized as the pre-registration operation for them as it ranks second in Sec. 4.1. In other words, we compare RFNet with the combination of NTG and SOTA fusion methods to evaluate the fusion performance and observe how significant the registration is for existing fusion methods. These fusion methods include DenseFuse [13], IFCNN [34], U2Fusion [30], PMGI [33] and MDLatLRR [14]. Besides, the fine registration of F2M is verified when the inputs suffer local misalignments.

Qualitative Results. Qualitative results on six typical unaligned image pairs are shown in Fig. 7. We analyze the results from three aspects. First, our method can register multi-modal images well as well as fuse their complementary information. As shown in the first two examples, the registration method fails to completely eliminate the parallaxes in two source images. The misalignments remain in the fusion results and result in disorganized scene content. By comparison, the joint coarse-to-fine registration in our method and the feedback of image fusion help correction the misalignments and improve the fusion performance. Second, our method can remove the overlapping shadows to present clear textures. As shown in the third and fourth rows, the slightly deficient registration accuracy causes the overlapping shadows and blurs the fusion results. By comparison, our method can finely remove the overlapping shadows and preserve more sharp edges. Third, our fusion results exhibit the most abundant and natural textures. In the last two examples, NIR images contain richer content than corresponding VIS images. In competitors, the blurred texture details in the VIS images affect the clarity of fusion results more or less. And in the fourth row, the trees in the result of IFCNN are closer to those in the NIR image than natural ones. By comparison, our results are suitable for the human visual perception system.

Quantitative Evaluation. We perform the quantitative evaluation of image fusion from two aspects. First, we assess the characteristics of fused images with average gradient (AG) [6], entropy (EN) [22] and standard deviation (STD). Second, we measure the similarity between the fused image and two source images with PSNR. It is worth noting that if the source images are unaligned, the fused images will suffer misalignments while the quantitative results may show fake improvements (*e.g.*, average gradients). To avoid the negative influence of this situation, we selected 35 image pairs that do not have noticeable misalignments after being processed by NTG/coarse registration module. The results are reported in Tab. 2. Our optimal results on AG, EN and STD show that our results contain the most texture details, the most amount of information and the most obvious contrast, respectively. Besides, our optimal result on PSNR indicates that the proposed fusion method produces least distortion and our fused images are closest to the source images.

External Verification on Object Detection. To evaluate the practical benefits of image fusion and its improved performances, an external verification is further performed. We compare the detection results by using YOLOv5[2] as the detector. As shown in Fig. 8, we perform the detections both on an unaligned image pair to validate the effect of registration accuracy and on an aligned image pair to validate the effect of fusion performance. In the first example, the misalignments in fusion results negatively impact the detections of the cars. When the images are well registered, the merged information from two modalities plays a positive role in promoting the detection result, as shown in the detection result on our registered and fused image. In the second example, the images are aligned. In this case, other fusion methods reduce the accuracy of detecting the stop sign compared with that in the VIS image. By comparison, our method increases the detection accuracy by fusing the information in the NIR image.

4.3. Ablation Study

Essential Factors in Coarse Registration Module. The essential factors in this module lie in three aspects, includ-

[2] https://github.com/ultralytics/yolov5

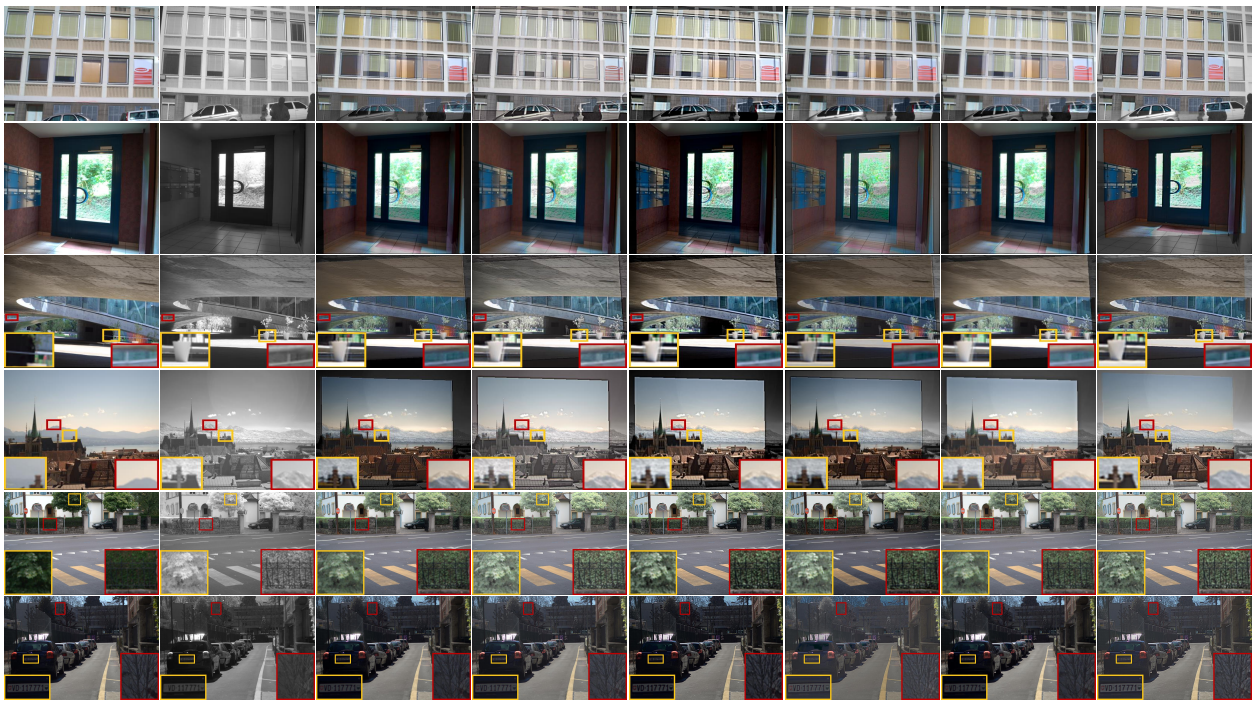


Figure 7. Fusion results on six unaligned multi-modal pairs.

Table 2. Quantitative comparisons of fusion performance (mean and standard variation, red: best, blue: second best, cyan: third best).

Methods	DenseFuse [13]	IFCNN [34]	U2Fusion [30]	PMGI [33]	MDLatLRR [14]	RFNet
AG	6.565±1.677	7.859±1.816	8.862±2.206	6.906±1.489	6.122±1.421	10.217±2.516
EN	7.092±0.372	7.290±0.295	7.105±0.403	7.104±0.324	7.222±0.276	7.325±0.288
STD	9.118±0.902	9.575±0.857	9.792±0.858	9.536±0.772	9.616±0.839	10.064±0.877
PSNR	64.678±2.011	65.359±2.376	64.641±1.666	63.712±1.729	66.269±2.334	66.363±2.150

Figure 8. Detection results on source images and different registration and fusion. The detector is YOLOv5.

ing image translation, network architecture of AffineNet and metric measuring the registration accuracy. We design three comparison experiments to separately validate their effectivenesses. The registration accuracy is uniformly evaluated by the NCC loss. **i)** We change the input of AffineNet and the loss is defined according to the inputs. We separately feed the descriptors defined in SCB [3], $\{I_x, I_y\}$ without translation, and $\{I_{x\to y}, I_y\}$ generated by our TransNet. The changes in losses shown in Fig. 9 demonstrate that the image-level inputs outperform the sparse descriptors. And the same-domain inputs further promote the convergence speed and performance. **ii)** The deformable convolution layers in AffineNet are replaced with traditional ones while traditional ones lead to a gradient explosion. **iii)** We compare the effect of NCC/L_1/L_2 loss as the metric. L_2 loss encounters a gradient explosion and Fig. 9 shows that NCC loss is superior to L_1 loss.

Fine Registration Performance of F2M. To validate the effectiveness of the fine registration in F2M on eliminating local misalignments, we perform two experiments by comparing F2M with two different competitors. One situation is that there are merely local parallaxes in source

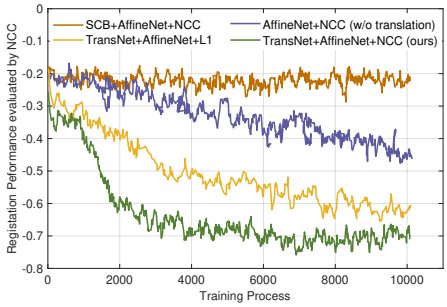
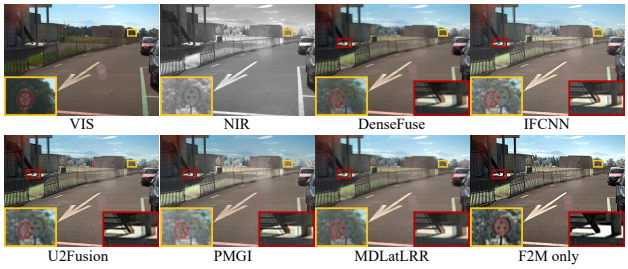
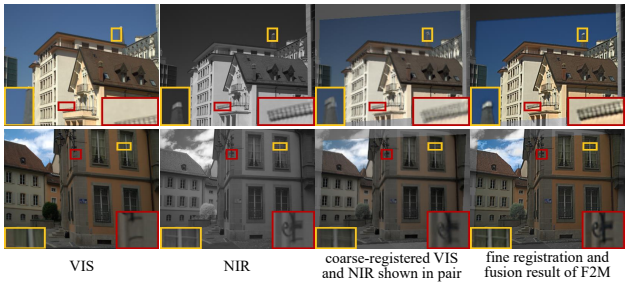


Figure 9. Changes in registration losses during training process.



(a) Dealing with images with small parallaxes without pre-registration



(b) Comparing the registration accuracy before and after F2M

Figure 10. Qualitative results to validate the effectiveness of the fine registration in F2M through two experiments.

images. In this case, existing registration methods or our coarse registration module are not performed in advance. Instead, we directly apply the SOTA fusion methods and F2M to deal with the unaligned images. As shown in Fig. 10(a), our F2M successfully removes the misalignments while they are still distinguishable in the results of the SOTA fusion methods. From the other side, we validate the fine-registration effect of F2M on the basis of coarse-registration results. As the coarse registration module does not have the fusion function, we show the coarse-registered VIS and NIR images in pair by the average weighted strategy rather than fused images. As shown in Fig. 10(b), the fine registration function of F2M helps remove the overlapping shadows in the coarse registration results.

Texture Preservation Strategies. We adapt the gradient channel attention mechanism, introduce the gradient loss, and set γ to a relatively high value to preserve texture details. To validate their effectiveness, we remove the attention mechanism, remove the gradient loss ($\delta = 0$) and set

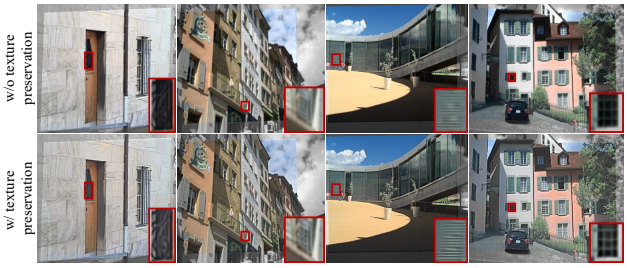


Figure 11. Qualitative comparison results with (w/) and without (w/o) texture preservation strategies.

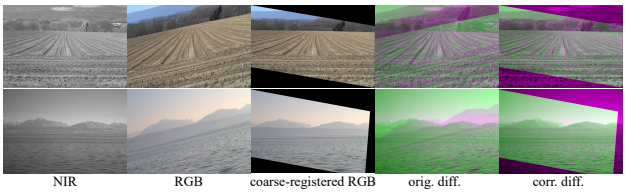


Figure 12. Some examples of scenarios where the performance of the coarse registration module is prone to degrade.

$\gamma = 0.5$. Comparison results are shown in Fig. 11. The results with texture preservation exhibit more texture details than those without using these strategies.

Limitations. It is typically difficult to establish strict correspondences between multi-modal images. In some cases, the scenes may show obvious cross-modal structure differences, such as field and forest (*e.g.*, the first example of Fig. 12). The image translation mainly adjusts the intensity, but rarely changes the scene content or structures (few edges are generated or eliminated). In other words, it is difficult for image translation to reduce the cross-modal structural differences. Moreover, in some other cases, the scenes may lack salient structures, such as water (*e.g.*, the second example of Fig. 12). These factors bring challenges to the coarse registration module which is based on image translation and NCC loss. Thus, in these cases, the registration accuracy of the coarse registration module is prone to degrade, as shown in the last column of Fig. 12.

5. Conclusion

In this paper, a new unsupervised multi-modal image registration and fusion method is proposed by mutually reinforcing the two individual tasks. The registration is handled in a coarse-to-fine approach. The coarse registration is modeled as an affine transformation and realized through an deformable convolution-based network and an image translation-based image-level loss function. The fine registration relies on the feedback of fusion. The fine-registered results further improve the fusion results. Also, we focus on texture preservation for both the feedback of fusion and image fusion itself. Experiments validate the effectiveness of the proposed method and mutually reinforcing framework.

References

[1] Moab Arar, Yiftach Ginger, Dov Danon, Amit H Bermano, and Daniel Cohen-Or. Unsupervised multi-modal image registration via geometry preserving image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13410–13419, 2020. 2, 4, 5, 6

[2] Matthew Brown and Sabine Süsstrunk. Multi-spectral sift for scene category recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 177–184, 2011. 5

[3] Si-Yuan Cao, Hui-Liang Shen, Shu-Jie Chen, and Chunguang Li. Boosting structure consistency for multispectral and multimodal image registration. *IEEE Transactions on Image Processing*, 29:5147–5162, 2020. 2, 5, 6, 7

[4] Chen Chen, Yeqing Li, Wei Liu, and Junzhou Huang. Sirf: Simultaneous satellite image registration and fusion in a unified framework. *IEEE Transactions on Image Processing*, 24(11):4213–4224, 2015. 2

[5] Shu-Jie Chen, Hui-Liang Shen, Chunguang Li, and John H Xin. Normalized total gradient: a new measure for multispectral image registration. *IEEE Transactions on Image Processing*, 27(3):1297–1310, 2017. 2, 5, 6

[6] Guangmang Cui, Huajun Feng, Zhihai Xu, Qi Li, and Yueting Chen. Detail preserved fusion of visible and infrared images using regional saliency extraction and multi-scale image decomposition. *Optics Communications*, 341:199–209, 2015. 6

[7] Yu Fu, Xiao-Jun Wu, and Tariq Durrani. Image fusion based on generative adversarial network consistent with perception. *Information Fusion*, 72:110–125, 2021. 2

[8] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1501–1510, 2017. 3

[9] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 448–456, 2015. 5

[10] Seungryong Kim, Dongbo Min, Bumsub Ham, Minh N Do, and Kwanghoon Sohn. Dasc: Robust dense descriptor for multi-modal and multi-spectral correspondence estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(9):1712–1729, 2017. 2, 5, 6

[11] Seungryong Kim, Dongbo Min, Bumsub Ham, Seungchul Ryu, Minh N Do, and Kwanghoon Sohn. Dasc: Dense adaptive self-correlation descriptor for multi-modal and multi-spectral correspondence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2103–2112, 2015. 2, 5, 6

[12] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[13] Hui Li and Xiao-Jun Wu. Densefuse: A fusion approach to infrared and visible images. *IEEE Transactions on Image Processing*, 28(5):2614–2623, 2018. 2, 6, 7

[14] Hui Li, Xiao-Jun Wu, and Josef Kittler. Mdlatlrr: A novel decomposition method for infrared and visible image fusion. *IEEE Transactions on Image Processing*, 29:4733–4746, 2020. 6, 7

[15] Yu Liu, Xun Chen, Juan Cheng, and Hu Peng. A medical image fusion method based on convolutional neural networks. In *Proceedings of the International Conference on Information Fusion (Fusion)*, pages 1–7, 2017. 2

[16] Yu Liu, Shuping Liu, and Zengfu Wang. A general framework for image fusion based on multi-scale transform and sparse representation. *Information Fusion*, 24:147–164, 2015. 2

[17] Jiayi Ma, Pengwei Liang, Wei Yu, Chen Chen, Xiaojie Guo, Jia Wu, and Junjun Jiang. Infrared and visible image fusion via detail preserving adversarial learning. *Information Fusion*, 54:85–98, 2020. 2

[18] Jiayi Ma, Han Xu, Junjun Jiang, Xiaoguang Mei, and Xiao-Ping Zhang. Ddcgan: A dual-discriminator conditional generative adversarial network for multi-resolution image fusion. *IEEE Transactions on Image Processing*, 29:4980–4995, 2020. 2

[19] Jiayi Ma, Wei Yu, Pengwei Liang, Chang Li, and Junjun Jiang. Fusiongan: A generative adversarial network for infrared and visible image fusion. *Information Fusion*, 48:11–26, 2019. 2

[20] Dwarikanath Mahapatra, Bhavna Antony, Suman Sedai, and Rahil Garnavi. Deformable medical image registration using generative adversarial networks. In *Proceedings of the IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 1449–1453, 2018. 2

[21] Fanjie Meng, Baolong Guo, Miao Song, and Xu Zhang. Image fusion with saliency map and interest points. *Neurocomputing*, 177:1–8, 2016. 2

[22] J Wesley Roberts, Jan A Van Aardt, and Fethi Babikker Ahmed. Assessment of image fusion procedures using entropy, image quality, and multispectral classification. *Journal of Applied Remote Sensing*, 2(1):023522, 2008. 6

[23] Colin Studholme, Corina Drapaca, Bistra Iordanova, and Valerie Cardenas. Deformation-based mapping of volume change from serial brain mri in the presence of local tissue contrast change. *IEEE Transactions on Medical Imaging*, 25(5):626–639, 2006. 2

[24] Carlo Tomasi and Roberto Manduchi. Bilateral filtering for gray and color images. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 839–846, 1998. 4

[25] Paul Viola and William M Wells III. Alignment by maximization of mutual information. *International Journal of Computer Vision*, 24(2):137–154, 1997. 5, 6

[26] Christian Wachinger and Nassir Navab. Structural image representation for image registration. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition-Workshops*, pages 23–30, 2010. 2

[27] Chengjia Wang, Giorgos Papanastasiou, Agisilaos Chartsias, Grzegorz Jacenkow, Sotirios A Tsaftaris, and Heye Zhang. Fire: unsupervised bi-directional inter-modality registration using deep networks. *arXiv preprint arXiv:1907.05062*, 2019. 2

[28] Qi Wei, José Bioucas-Dias, Nicolas Dobigeon, and Jean-Yves Tourneret. Hyperspectral and multispectral image fusion based on a sparse representation. *IEEE Transactions on Geoscience and Remote Sensing*, 53(7):3658–3668, 2015. 2

[29] Qi Xie, Minghao Zhou, Qian Zhao, Deyu Meng, Wangmeng Zuo, and Zongben Xu. Multispectral and hyperspectral image fusion by ms/hs fusion net. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1585–1594, 2019. 2

[30] Han Xu, Jiayi Ma, Junjun Jiang, Xiaojie Guo, and Haibin Ling. U2fusion: A unified unsupervised image fusion network. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(1):502–518, 2022. 6, 7

[31] Yong Yang, Jiaxiang Liu, Shuying Huang, Weiguo Wan, Wenying Wen, and Juwei Guan. Infrared and visible image fusion via texture conditional generative adversarial network. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(12):4771–4783, 2021. 2

[32] Hao Zhang, Zhuliang Le, Zhenfeng Shao, Han Xu, and Jiayi Ma. Mff-gan: An unsupervised generative adversarial network with adaptive and gradient joint constraints for multi-focus image fusion. *Information Fusion*, 66:40–53, 2021. 2

[33] Hao Zhang, Han Xu, Yang Xiao, Xiaojie Guo, and Jiayi Ma. Rethinking the image fusion: A fast unified image fusion network based on proportional maintenance of gradient and intensity. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12797–12804, 2020. 6, 7

[34] Yu Zhang, Yu Liu, Peng Sun, Han Yan, Xiaolin Zhao, and Li Zhang. Ifcnn: A general image fusion framework based on convolutional neural network. *Information Fusion*, 54:99–118, 2020. 6, 7