IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received July 30, 2019, accepted August 7, 2019, date of publication September 2, 2019, date of current version September 26, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2938858

Towards Accurate and Robust Multi-Modal Medical Image Registration Using Contrastive Metric Learning

JINRONG HU[1,4], SHANHUI SUN[2], XIAODONG YANG[1], SHUANG ZHOU[3], XIN WANG[2], YING FU[1], JILIU ZHOU[1], YOUBING YIN[2], KUNLIN CAO[2], QI SONG[2], AND XI WU[1]

[1]Department of Computer Science, Chengdu University of Information Technology, Chengdu 610225, China
[2]CuraCloud Corporation, Seattle, WA 98004, USA
[3]Philips Research China, Shanghai 210000, China
[4]Department of Computer and Soft Engineering, Xihua University, Chengdu 610039, China

Corresponding authors: Shanhui Sun (shanhuis@curacloudcorp.com) and Xi Wu (xi.wu@cuit.edu.cn)

This work was supported in part by the National Science Foundation of China under Grant 61303126 and Grant 61602390, in part by the Ministry of Education, China: Chunhui Project under Grant Z2015108, in part by the Sichuan Science and Technology Program under Grant 2016RZ0051 and Grant 2018RZ0072, and in part by the Open Research Fund from Province Key Laboratory of Xihua University under Grant szjj2013-022.

ABSTRACT Multi-modal medical image registration takes an essential role in image-based clinical diagnosis and surgical planning. It is not trivial due to appearance variations across different modalities. Rigidly aligning two images is used to register rigid body structure, and it is also usually the first step for deformable registration with a large discrepancy. In the field of computer vision, one well-established method for image alignment is to find corresponding points from two images, and image alignment is based on identified corresponding points. Our method lies in this category. Feature representation is crucial in finding corresponding points. However, conventional feature representation like SIFT does not take multi-modal information into account, and thus, it fails. In this paper, we propose a Convolution Neural Network Feature-based Registration (CNNFR) method for aligning the multi-modal medical image. The important component in this method is learning keypoint descriptors using contrastive metric learning, which minimizes the difference between two feature representations from two corresponding points and maximizes difference of two feature representation from two distant points. Also, a transfer learning-based CNNRF (TrCNNRF) is proposed to improve the generalization learning performance when the training data are insufficient. Experimental results demonstrate that the proposed methods can achieve superior performance regarding both accuracy and robustness, which can be used to rigidly register multi-modal images and provide an initial estimation for non-rigid registration in clinical practices.

INDEX TERMS Medical image registration, convolution neural networks (CNNs), transfer learning, feature descriptor, contrastive metric learning.

I. INTRODUCTION

Medical imaging provides insights into the size, shape, and spatial relationships among anatomical structures. For instance, CT is handy for skeletal structures and dense tissue, whereas MRI provides a view of soft tissue. Aligning these different modalities can provide useful complementary information for more efficient cancer detection, disease diagnosis, and treatment planning.

The associate editor coordinating the review of this manuscript and approving it for publication was Ruqiang Yan.

There is abundant literature on the problem of multi-modal medical image registration ([1]–[15]). The goal of image registration is finding an optimal transformation to completely align the fixed and moving images together into one coordinate system. However, due in no small appearance discrepancy across different modalities (i.e., CT and MR), robust and fast multi-modal image registration is still not fully solved.

The methods of multi-modal medical image registration can be classified into rigid and deformable. Specifically, rigid registration is often used in the alignment of rigid body

structures; and it is also usually the first step for deformable registration with a large displacement or discrepancy. Therefore, robust, accurate and fast rigid alignment algorithms are still highly required.

The rigid registration methods usually fall into two categories: intensity-based [16]–[19] and feature-based. In practice application, the feature-based algorithms play a vital role in medical image registration because of their computational efficiency and robustness. Concretely, the keypoint feature-based methods are widely employed since the keypoints are the most commonly used features in clinical applications, such as the image-guided surgery and radiation systems.

The keypoint feature-based approaches estimate the geometric transform parameters and find corresponding keypoints from the fixed and moving images. The correspondence of the keypoints in the fixed and moving images can be built via matching the descriptors of keypoints. However, the problem of deciding if two keypoints correspond to each other or not accurately and robustly is quite challenging as the great variability of tissue or organ appearance in the multi-modal medical images, which results in the lack of a general rule to establish keypoint correspondence.

Previous works [20]–[26] used hand-designed feature descriptors like the Scale Invariant Feature Transform (SIFT) and its variants to address keypoint matching problem. In the SIFT algorithm, the local extrema in a difference of Gaussian scale space are selected as a keypoint. A 128-dimension descriptor is generated for the keypoint using the gradient magnitude and orientation in a local neighborhood of the keypoint. Due to the gradient direction variation caused by the vast intensity difference in multi-modal medical images, SIFT tends to build dissimilar descriptors at corresponding keypoints leading to wrong matching points. As a result, it cannot meet the accuracy and robustness requirement of multi-modal medical image registration.

To overcome this problem, some SIFT-related methods have been adapted to align multi-modal natural images [22]–[24], [26], [27] or remote sensing images [25], [28]. In [24], Hossein-Nejad and Nasri proposed the Magnitudes and Occurrences of Gradient SIFT (MOG-SIFT), which utilizes the gradient magnitude, gradient occurrence and gradient orientation information to build feature descriptor. In [25], Ye *et al.* proposed the Position Scale Orientation SIFT (PSO-SIFT), which utilizes a new gradient definition and a feature matching method by combining the position, scale, and orientation of each keypoint to improve the matching accuracy. These methods promoted the invariant and discriminative abilities of SIFT descriptors by attaching a wide diversity of gradient information. However, they can not lend the descriptor to change in appearance variations between different modalities of medical images. Fig. 1 shows a case using the SIFT descriptors to match keypoints between CT and MR. Because the keypoints in CT and MR may have different context gradient patterns, the SIFT fails to find the corresponding points.

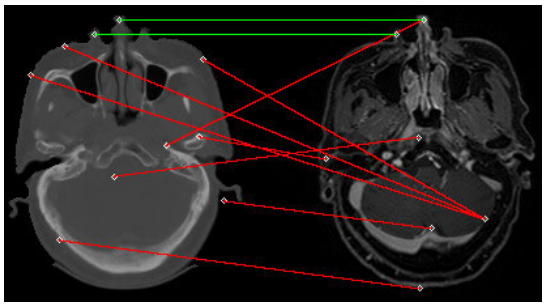


FIGURE 1. A failed matching example of SIFT method for CT and MR image pair. The left one is CT, the right one is MR. And the green line means correct matching, the red line means error.

Unlike SIFT descriptor based on the image's gradient information, MP Heinrich et al. proposed a modality independent neighborhood descriptor (MIND) to extract distinctive structure in a local neighborhood [29]. The MIND is under the assumption that the intensity distribution of an anatomical structure corresponds across modalities, and it is robust to the most considerable differences between modalities. However, due to the great variability of tissue or organ appearance caused by different physical principles of medical imaging, which leads to lack of correspondence on particular intensity distribution across different modalities medical images. As a result, the MIND can't provide a good feature representation of the keypoint's local neighborhood and fails to find the corresponding points.

In our work, we proposed a keypoint matching-based registration framework for aligning multi-modal images. Our key component in the framework is a new approach to learn representative descriptors of keypoints for multi-modal images utilizing contrastive metric learning to solve above problem. The utilized contrastive metric learning in the context of the deep neural network was first studied in the literature [30], where they coined the method as Siamese network.

The idea of multi-modal descriptors learning is that the learned descriptor has a small Euclidean distance in the feature space for two corresponding keypoints and a large Euclidean distance for non-corresponding keypoints. We coined our method as CNNFR for learning network weights based on positive and negative CT and MR patch pairs from nasopharyngeal carcinoma (NPC) patient data. Also, to minimize generalization error, we pre-train the network on natural images from UBC dataset [31]. We coined the pretrain-based method as TrCNNFR. Unlike any other deep learning-based one-step transformation estimation techniques [15], we use the learned descriptor to augment the performance of keypoint feature-based registration. To the best of our knowledge, this is the first work that employs contrastive metric learning to solve the problem of cross modalities medical image registration. The validation demonstrates that the proposed method achieves state-of-the-art registration accuracy.

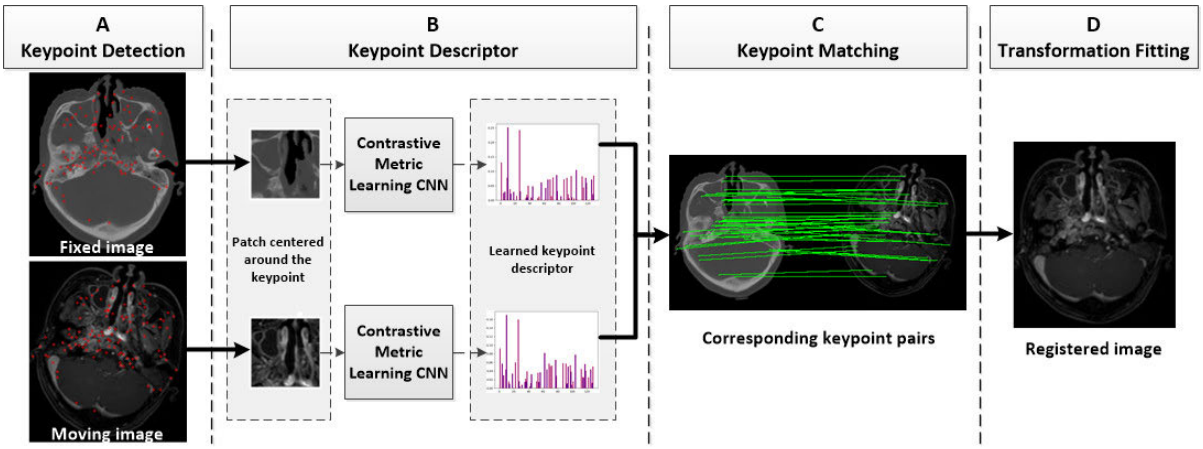


FIGURE 2. Workflow of the proposed CNNFR technique. A is the keypoint detection phase. We use the DoG maxima to detect the keypoints. B is the keypoint descriptor stage. We get patches centered around the keypoints at first. Then we employ the Contrastive Metric Learning CNN to extract the keypoint descriptor from its located patch. C is the keypoint matching step. D is the transformation fitting stage. At last, the registered image is gained by using the transformation to warp moving image.

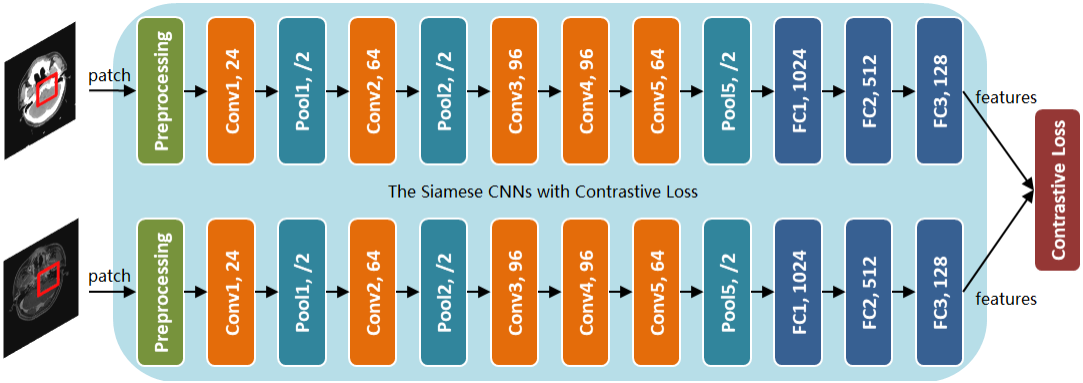


FIGURE 3. The detailed architecture of the Siamese network, which consists of two CNNs with identical parameters.

II. METHODS

A. REGISTRATION ALGORITHM

Fig. 2 illustrates the overview of the proposed method: Convolution Neural Network Feature-based Registration (CNNFR). It consists of four main components: keypoint detection, keypoint descriptor, keypoint matching, and transformation fitting. We describe each component in the following parts.

1) KEYPOINT DETECTION

The keypoints are those salient points in the images such as corner points. In this step, we use the Difference of Gaussian (DOG) maxima to detect the keypoints.

2) KEYPOINT DESCRIPTOR

The keypoint descriptor is the local features of point. A good descriptor can significantly improve the keypoint matching performance. In this stage, without resorting to manually-designed features, we propose to extract stable, distinctive and uniformly distributed features for keypoints in cross modality medical image with contrastive metric learning. Specifically, we learn the keypoint descriptor based on the Siamese network.

Fig. 3 presents the structure of the Siamese network, which consists of two CNNs with identical weights. In each CNN, the first layer is a preprocessing layer, which normalizes the intensity value $m \in [0, 255]$ of each pixel in the input patch centered at keypoint to $\frac{m-128}{160}$, the last layer outputs the features. The parameters of each layer are summarized in Table 1. To reduce the risk of overfitting caused by very deep networks (This is well documented in [32]), we don't use the very deep VGG network.

To ensure not only that the loss of the Siamese network for a pair of patches from the two corresponding keypoints is low, but also that the loss for a pair from non-corresponding keypoints is large, we design the Siamese network to minimize the Contrastive loss defined as follows:

$$L = \frac{1}{2N}\sum_{i=1}^{N} y_i d_i^2 + (1 - y_i)max(margin - d_i, 0)^2 \quad (1)$$

where $d_i = \|x_{i1} - x_{i2}\|_2$ is the Euclidean distance between the intensity patch x_1 and x_2, y_i is the binary label for input

TABLE 1. Network parameters. All convolution and fullconvolution layers use ReLU activation except for FC3.

Name	Type	Output dim	Patch size	stride
Conv1	Convolution	$64 \times 64 \times 24$	7×7	1
Pool1	MaxPooling	$32 \times 32 \times 24$	3×3	2
Conv2	Convolution	$32 \times 32 \times 64$	5×5	1
Pool2	MaxPooling	$16 \times 16 \times 64$	3×3	2
Conv3	Convolution	$16 \times 16 \times 96$	3×3	1
Conv4	Convolution	$16 \times 16 \times 96$	3×3	1
Conv5	Convolution	$16 \times 16 \times 64$	3×3	1
Pool5	MaxPooling	$8 \times 8 \times 64$	3×3	2
FC1	FullConvolution	1024	-	-
FC2	FullConvolution	512	-	-
FC3	FullConvolution	128	-	-

pairs, where 1 indicates match and 0 otherwise. *margin* is a constant with a default value of 1 in Caffe [33].

Moreover, to make the learned descriptor more robust to the variety of characteristics in different modalities, we train the Siamese network with Contrastive loss based on numerous patch pairs centered at the corresponding and non-corresponding keypoints in CT and MR pairs from the clinical NPC patient described in Section III-A.

The learned keypoint descriptor extracted by our Siamese network can significantly outperform handed crafted feature descriptor.

3) KEYPOINT MATCHING

In this phase, we identify a set of the corresponding keypoint pairs by matching the keypoint descriptors. Given S_1 represents the set of keypoint descriptors in the fixed image, S_2 represents the set of keypoint descriptors in the moving image, and $Dis(p, q)$ represents the Euclidean distance between the feature descriptor p and q. Then we can sort the members of S_2 by their distance to a descriptor in S_1. Let $p \subset S_1$, $q_i \subset S_2$ be the i-th nearest member of S_2 to p, the matching score is defined as:

$$MS(p, S_2) = \frac{Dis(p, q_1)}{Dis(p, q_2)} \tag{2}$$

$MS(p, S_2)$ is small when the match between p and q_1 is particularly distinctive, i.e., p is much closer to q_1 than any other member of S_2. Thus, we say that p matches q_1 if $MS(p, S_2)$ is below some threshold η, which is also called as matching ratio. This prevents matching when keypoints are locally similar, which often occurs in medical images.

4) TRANSFORMATION FITTING

In this process, we solve an affine geometric transform according to the obtained correspondences. Given a coordinate $o \in R^n$ with parameters $A \in R^{n \times n}$ and $b \in R^n$, an affine transform has the form:

$$o' = Ao + b. \tag{3}$$

We fit the affine transform by the least square, requiring at least $n + 1$ matches for uniqueness. Some of the matches will be erroneous, so we reject outliers by Random Sample Consensus (RANSAC) as described in [34]. This attempts to find the transformation with the most inliers, by iteratively fitting transforms to subsets of the data. The final transformation is the least-squares fitting the inliers.

B. TRANSFER LEARNING-BASED CNNFR (TRCNNFR)

It is not easy to obtain a large number of annotated medical image pairs because of challenging data acquisition and a great number of annotation efforts. Although we could augment a large number of image patches, we are not able to create a new distribution from them. However, low-level network representations are usually common across different applications such as edge, corners, etc. which we could leverage from other datasets. In this work, we utilized transfer learning from natural images. Specifically, we train our network on the UBC dataset, which is a collection of patches extracted around real interest points from several internet photo collections published in [31]. Next, we fine-tuned the network using the CT and MR pairs described in Section III-A.

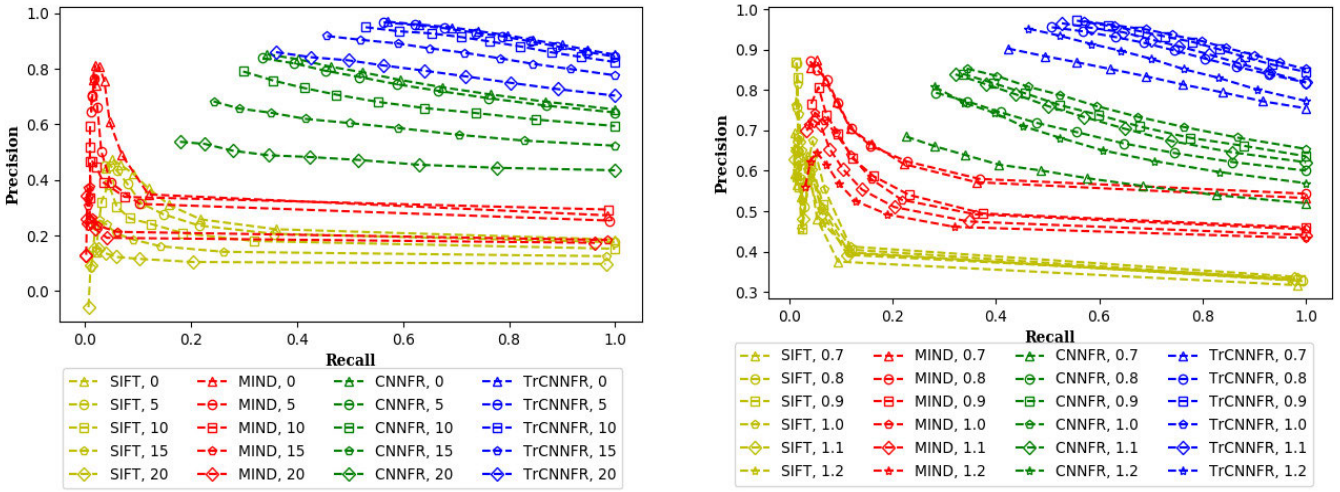
III. EXPERIMENT AND RESULTS

A. DATA AND GROUND TRUTH

This study has been conducted using CT and MR images of 100 Nasopharyngeal Cancer (NPC) patients (male/female: 52/48; mean age $\pm$ standard deviation: 50.3 $\pm$ 11.2 years; age range: 21-76 years old) underwent chemoradiotherapy or radiotherapy at West China Hospital. The CT images were obtained by a Siemens SOMATOM Definition AS+ system, with a voxel size ranges from $0.88 \times 0.88 \times 3.0\ mm^3$ to $0.97 \times 0.97 \times 3.0\ mm^3$. The MR images were captured by a Philips Achieva 3T scanner. We used in this study the T1-weighted images with contrast. The in-plane resolution is $0.61 \times 0.61\ mm^2$ and a slice spacing of 0.8 *mm*. It is noted that every scan contains nasopharyngeal cancer.

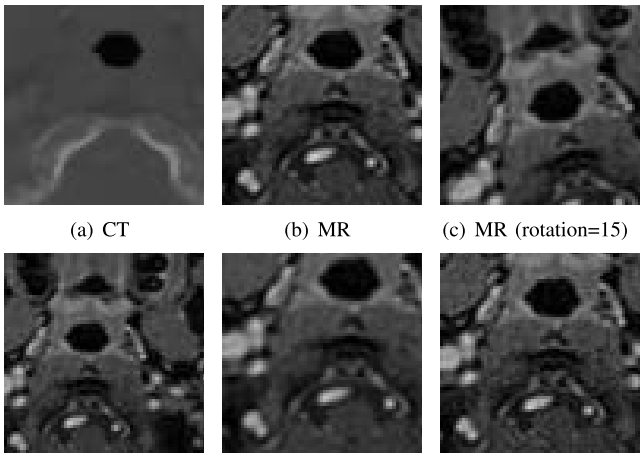
We resampled the images to have a voxel size of $1 \times 1 \times 1\ mm^3$. Because of the different imaging ranges covered by those images, we only kept the range from eyebrow and chin. Moreover, because the ground truth alignment of these two image modalities is unfortunately not easily obtainable, the standard of alignment is estimated using an off-the-shelf toolbox Elastix [17] for the sake of efficiency. It is noted that the alignment is performed in 3D using Elastix with standard parameters.

Then we randomly selected 70 registered CT and MR volumes and used them to train the proposed methods (Section III-B). For each volume pair, we randomly picked 15 paired CT and MR slices and augmented the data by rotating, scaling and adding noise. Given one CT or MR slice, two images were generated by adding random noise following Gaussian distributions $N(0, 5)$ and $N(0, 10)$ to the slice, respectively; eight images were obtained by rotating the slice by a degree from -20 to 20 with a step of 5; six images were acquired by scaling the slice with a factor in [0.7, 1.3] with a step of 0.1.



(a) Rotation (b) Scaling

FIGURE 4. Precision-recall curve comparison for the keypoint matching using SIFT, MIND, CNNFR and TRCNNFR descriptors with η in the interval [0.6, 1.0] with a step of 0.05.



(a) CT (b) MR (c) MR (rotation=15)

(d) MR (scaling=0.8) (e) MR (scaling=1.2) (f) MR (δ^2=10)

FIGURE 5. Some patches with positive label. All patches with size 64 × 64 were extracted from the identical keypoints in alignment CT and MR images and their respective augmented images. (a) and (b) are patches from paired CT and MR images correspondingly. (c) to (f) are patches from the augmented MR images with rotating −15-degree, scaling 0.8 and 0.2 times, and white Gaussian noise with variance 10 separately.

For each slice, we used the extrema in a Difference of Gaussian (DoG) scale space as the keypoints. Then we extracted image patches of size 64 × 64 centered at the keypoints. Meanwhile, we computed the corresponding points of the keypoints in the augmented images, and extracted patches of size 64 × 64 centered at the corresponding points. We used the patch pairs as training and validation data. A patch pair received a positive label if the patches were generated from the same slices or two corresponded slices, and their corresponding keypoints are entirely identical. Accordingly, it got a negative label if the patches were generated from the different slices or two non-corresponding slices, and the absolute distance between their keypoints is more than 50mm. Fig. 5 shows some patches with the positive label.

B. EXPERIMENT SETUP

We compared our proposed methods: CNNFR and TrCNNFR to SIFT, MIND, affine image registration network (AIRNet) [35], and the intensity-based registration method using 2D ELASTIX toolbox. We used 50 NPC patients' images for training, another 20 NPC patients' images for validation, and the rest 30 patients' images for testing. We generated image patch pairs in the way described in Section III-A. We used 500,000 and 200,000 patch pairs as training data and validation data for CNNFR accordingly, which are with the same proportion of positive and negative samples.

We implemented CNNFR and TrCNNFR using Caffe [33]. For CNNFR, the training batch size was set to 256. We used SGD optimizer with momentum. The learning rate started at 0.01, with momentum 0.9. The learning rate was reduced after each epoch by a factor of 0.96. The training process was terminated after 30,000 iterations. The training of each epoch took about 10 minutes on an NVIDIA Tesla K40C GPU. CNNFR was trained with 20 epochs, and it converged before the end of training. TrCNNFR was first trained on the UBC dataset, with a learning rate of 0.001, momentum 0.9, weight decay 0.0. SGD was employed as an optimizer, and the batch size was set to 256. The training process was terminated after 10,000 iterations. Then, the network was fine-tuned using the NPC dataset with the same learning setting, and the learning process was terminated after 30,000 iterations. Our codes for this work are available on *https://github.com/CUIT-MIA/CNNFR*.

For one keypoint in the moving image, it is considered as a "true correspondence" to its corresponding keypoint in the fixed image if the distance between its mapped point with the corresponding keypoint is within ϵ pixels. Here, ϵ was set to 3. The RANSAC parameter ω is used to determine if a matched point pair is inlier or outlier in transformation fitting procedure. In our experiments,

ω was equal to 2. We computed the local SIFT descriptors and MIND by using the open-source VLFeat toolbox on *http://www.vlfeat.org/* and the MIND code shared by MP Heinrich on *http://www.mpheinrich.de/software.html* accordingly. At the same time, we implemented the 2D AIRNet according to its original paper [35]. For the 2D Elastix parameters, we chose advanced mattes mutual information as the optimization criterion, and adaptive stochastic gradient descent as the optimization routine and *similarity transform* as the transformation model. Four image pyramids (resolutions) were used, each with 500 iterations.

C. KEYPOINTS MATCHING

In this section, we compared our proposed matching methods with SIFT and MIND matching algorithm. The descriptors are thus CNNFR, TrCNNFR, SIFT, and MIND. Let the true correspondence be the one within $\epsilon = 3$ mm of the ground truth, a positive be a matched keypoint in the fixed image, a false positive is a keypoint which was assigned an incorrect match, while a false negative is a keypoint for which a true correspondence exists in the other image, but is not assigned a match. From these definitions, we computed the standard precision and recall scores as follows:

$$Precision = \frac{true\ positives}{(true\ positives + false\ positives)} \tag{4}$$

$$Recall = \frac{true\ positives}{(true\ positives + false\ negatives)} \tag{5}$$

For each patient, along with the axial direction, we randomly chose three paired CT-MR images for validating descriptor performance in the matching algorithm concerning the precision-recall curve. Fig. 4 illustrates descriptor performance of SIFT, MIND, CNNFR and TrCNNFR for different registration tasks. Fig. 4(a) presents the precision-recall curves for registration task where the moving images were generated by rotating the ground truth by degrees from 0 to 20 with a step of 5. Fig. 4(b) presents results for registration task where the moving images were generated by scaling the ground truth by factors in [0.7, 1.3] with a step of 0.1. The precision and recall scores were gained at different matching thresholds η (as described in Section II-A.3) in the interval [0.6, 1.0] with a step of 0.05.

Specifically, the lower the matching threshold η is the higher the precision is. That is because the matching is more distinctive when the value of η is small, and thus less ''true correspondence'' can be identified. When $\eta = 0.6$, SIFT can barely find any positive matching and the denominator in Equation 4 equals 0. In this case, we set the precision to -1. So there is point whose coordinate value, the average precision score is less than zero in curve ''SIFT, 20'' in Fig. 4(a). As the value of η increases, false matching will exist. The precision decreases and the recall increases. When $\eta = 1$, any point can find a matching. In this case, the precision scores of TrCNNFR for all registration tasks are around 0.8, while the precision scores of SIFT are only around 0.2.

D. IMAGE REGISTRATION

In this section, we investigate the registration performance on the accuracy of six methods: intensity-based registration using ELASTIX toolbox, affine image registration network AIRNet, SIFT, MIND, CNNFR and TrCNNFR. The ELASTIX toolbox consists of a collection of algorithms that are commonly used to solve the rigid and non-rigid medical image registration problems. Due to its accessibility, usability and state-of-the-art results, it is used to compare different registration methods usually. AIRNet is a self-supervised and end-to-end deep learning registration method, which employs a CNN to predict the transformation parameters to register rigidly. It is considered a deep learning baseline for rigid registration. The registration accuracy is measured by Target Registration Error (TRE), which is the root-mean-square on the distance errors overall landmark pairs for each patient sample. Using the keypoints as landmarks, TRE is calculated as follows:

$$TRE = \sqrt{\frac{1}{M}\sum_{1}^{M} \|KP_i^{GT} - T \cdot KP_i\|_2^2} \tag{6}$$

where KP_i^{GT} denotes a keypoint in the ground truth moving image, KP_i is the keypoint in the moving image, T is the estimated transformation matrix after transformation fitting, M is the number of keypoints.

TABLE 2. TREs of the rotated registration task on image pairs from the testing NPC.

Rotation Degree	0	5	10	15	20
ELASTIX	0.90 ± 0.39	0.86 ± 0.37	0.92 ± 0.53	1.13 ± 0.84	1.42 ± 1.15
AIRNet	0.98 ± 0.57	1.11 ± 0.71	1.21 ± 0.76	1.48 ± 0.83	1.76 ± 0.83
SIFT	7.18 ± 4.35	8.30 ± 5.18	8.03 ± 4.01	10.79 ± 9.83	12.11 ± 7.01
MIND	5.57 ± 1.16	6.1 ± 1.28	6.84 ± 3.22	7.43 ± 4.33	9.95 ± 6.13
CNNFR	0.08 ± 0.64	0.13 ± 1.06	0.20 ± 1.06	0.16 ± 0.80	0.39 ± 1.48
TrCNNFR	0.02 ± 0.01	0.02 ± 0.01	0.02 ± 0.01	0.02 ± 0.01	0.02 ± 0.01

TABLE 3. TREs of the scaled registration task on image pairs from the testing NPC.

Scaling Factor	0.7	0.8	0.9	1.0	1.1	1.2
ELASTIX	1.08 ± 0.72	0.98 ± 0.53	0.89 ± 0.38	0.90 ± 0.39	0.84 ± 0.40	0.80 ± 0.43
AIRNet	2.88 ± 0.72	1.98 ± 0.53	1.49 ± 0.38	0.90 ± 0.39	1.84 ± 0.40	2.65 ± 0.43
SIFT	9.86 ± 5.71	10.86 ± 6.42	7.96 ± 4.32	7.18 ± 4.35	8.30 ± 6.03	8.68 ± 3.39
MIND	7.08 ± 0.72	6.28 ± 0.53	5.56 ± 0.38	4.90 ± 0.39	5.84 ± 0.40	6.80 ± 0.43
CNNFR	1.13 ± 2.24	0.43 ± 1.48	0.12 ± 0.64	0.08 ± 0.64	0.01 ± 0.77	0.08 ± 0.47
TrCNNFR	0.21 ± 0.86	0.01 ± 0.01	0.02 ± 0.01	0.02 ± 0.01	0.05 ± 0.30	0.01 ± 0.01

For each patient, along with the axial direction, we randomly chose three paired CT-MR images for validating registration algorithms. Table 2 demonstrates TREs of the six methods for registration tasks where the ground truth moving images were rotated clockwise by a certain degree. Table 3 presents the TREs of these methods for registration tasks where the ground truth moving images were scaled by a specific factor. As we can see from the tables, the TREs of

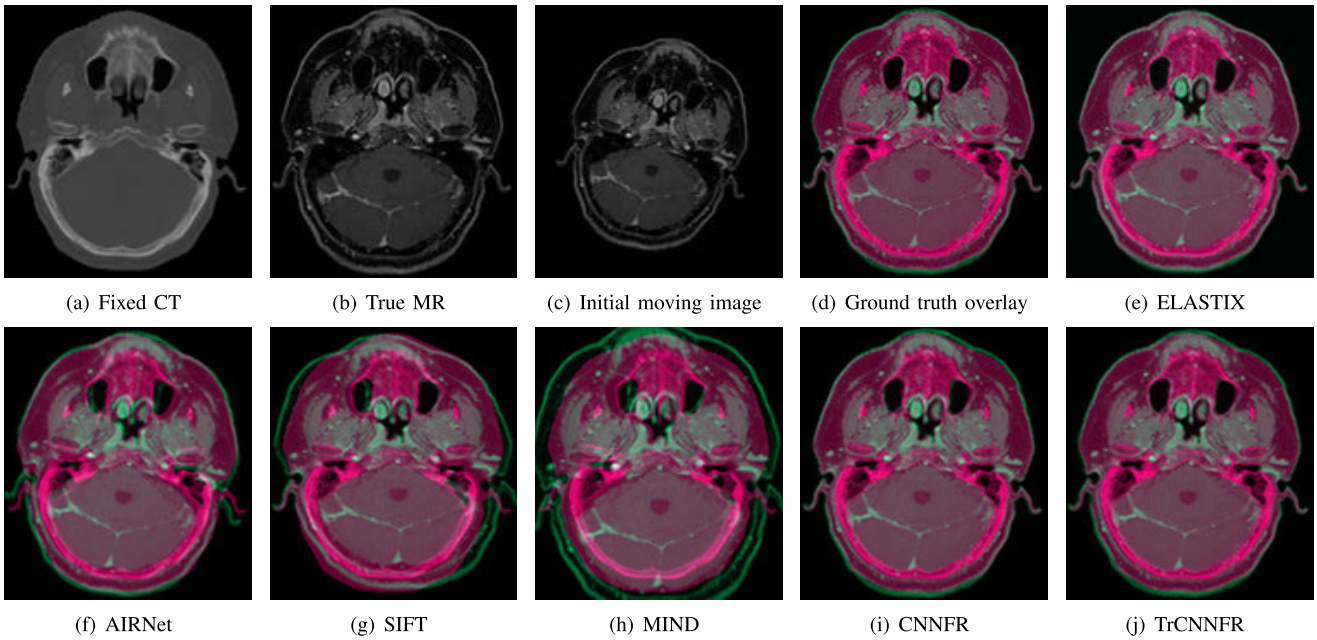


FIGURE 6. Illustration example of different registration methods. (a) Fixed CT. (b) Corresponding MR slice (ground truth). (c) Moving MR generated through rotating image (b) by 15° and scaling factor 0.8. (d) Ground truth MR image overlayed on the fixed image. (e) to (j) Registration results: ELASTIX, AIRNet, SIFT, MIND, CNNFR and TrCNNFR overlayed on the fixed CT.

CNNFR and TrCNNFR are much smaller than the TREs of AIRNet, SIFT and MIND for all cases. At the same time, the TREs of our methods are also smaller than the state-of-the-art intensity-based method (ELASTIX) in case of rotation and scaling.

TABLE 4. Time cost of ELASTIX, AIRNet, SIFT, MIND, CNNFR and TrCNNFR.

Methods	ELASTIX	AIRNet	SIFT	MIND	CNNFR	TrCNNFR
Time(in second)	43.9 ± 13.2	0.20 ± 0.01	1.1 ± 0.1	7.2 ± 0.9	1.5 ± 0.1	1.5 ± 0.1

To validate the computational efficiency of ELASTIX, AIRNet, SIFT, MIND, CNNFR and TrCNNFR, we calculated the average time costs of registering 90 pairs of 232×320 CT and MR slices for registration task mentioned above. Experiments are conducted on a PC with Linux OS Ubuntu 16.04, and Intel Xeon CPU E5-2640 v3 @ 2.6GHZ, and 128 GB of RAM, and with an NVIDIA Tesla K40C GPU. Table 4 displays the average time-consuming of the six methods. As we can see from Table 4, the running time of SIFT, CNNFR and TrCNNFR are around a second, and the running time of MIND is around seven seconds. However, the computational complexity of ELASTIX is very high; it is up to 43.9 seconds with std 13.2, which is about 29 times slower than TrCNNFR. The end-to-end deep learning registration method AIRNet is very computationally efficient; it is average running time only 0.2 seconds.

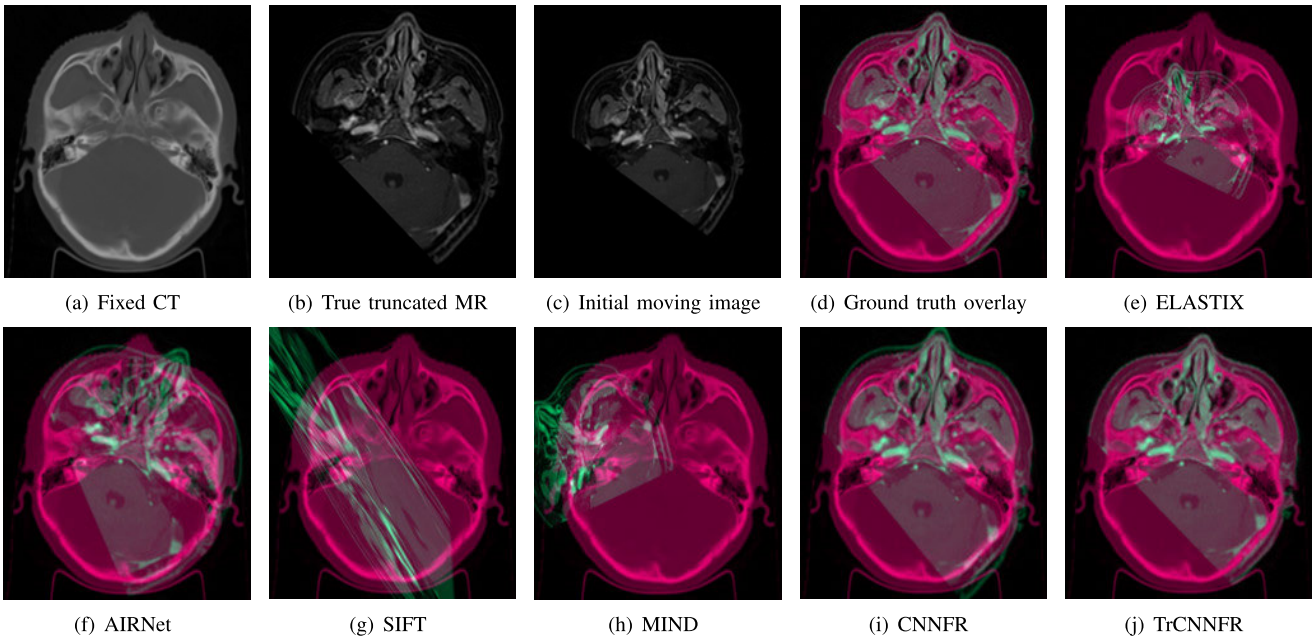
Fig. 6 illustrates one example in the test NPC data set of registration results of different methods: ELASTIX, AIRNet, SIFT, MIND, CNNFR and TrCNNFR. In this example, the TREs of the six methods are 1.11, 2.06, 4.13, 4.95, 1.12 and 0.95, respectively. The proposed TrCNNFR achieves the lowest TRE.

In Fig. 7, we showed one example of how well the different registration algorithms are handling missing data in the moving image. In this example, the TREs of ELASTIX, AIRNet, SIFT, MIND, CNNFR and TrCNNFR are 5.44, 4.37, 11.07, 10.26, 2.11 and 0.01, respectively. Fig. 8 illustrates an example of how well the different registration algorithms handling a low overlap between fix and moving images. In this example, the TREs of ELASTIX, AIRNet, SIFT, MIND, CNNFR and TrCNNFR are 4.77, 3.86, 44.29, 37.93, 1.34 and 0.02, respectively. In both cases, SIFT, AIRNet, MIND and ELASTIX are not able to register the pair images.

E. UNSEEN BODY PARTS AND IMAGE MODALITIES

To demonstrate the generalization capability of our proposed method, we applied our methods into two new tasks (unseen applications). The first task is the unseen body part CT-MRI registration. The second one is the unseen image modality image registration.

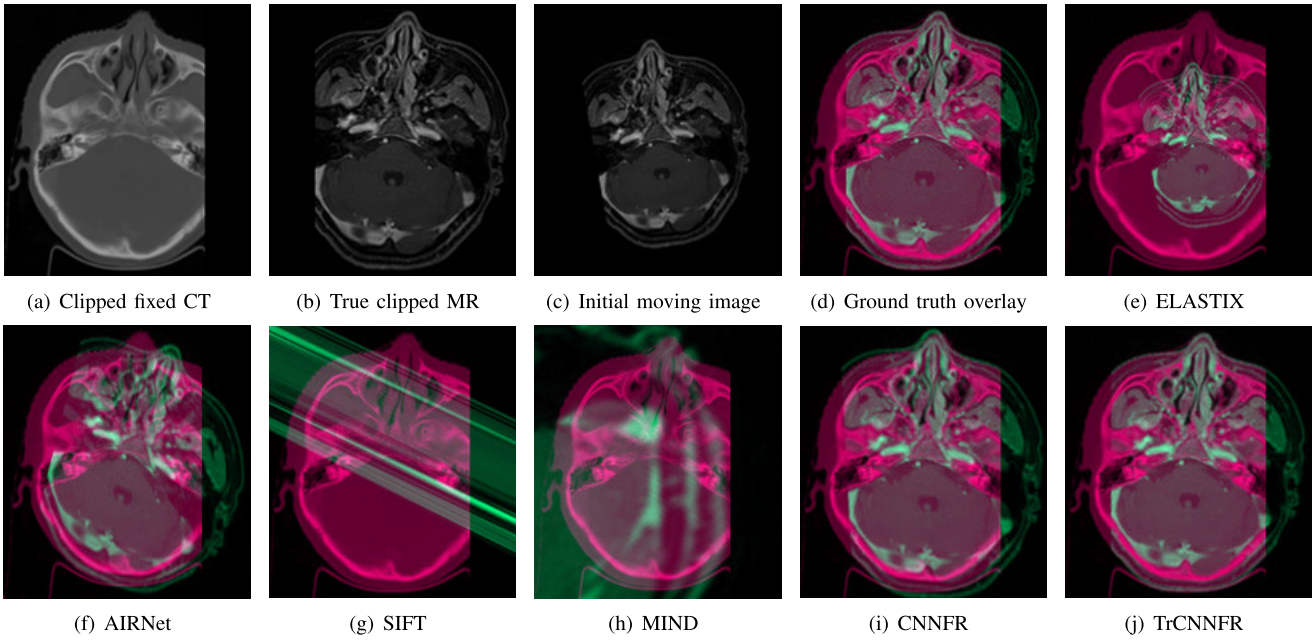
For the first task, the Siamese network was trained on the image patches taken from the CT-MR slices between eyebrow and chin (testing NPC), while the testing data were CT-MR slices between chin and shoulder (unseen body parts). For each NPC patient, we randomly selected ten pairs of CT-MR slices between chin and shoulder. Table 5 presents TREs of the six methods when the ground truth moving images were rotated clockwise by a certain degree. Table 6 provides the TREs of these methods for when a particular factor scaled the ground truth moving images.



(a) Fixed CT (b) True truncated MR (c) Initial moving image (d) Ground truth overlay (e) ELASTIX

(f) AIRNet (g) SIFT (h) MIND (i) CNNFR (j) TrCNNFR

FIGURE 7. Illustration example of different registration methods for missing data. (a) Fixed CT. (b) Corresponding MR slice (ground truth) with missing data perturbation. (c) Initial moving image generated through rotating image (b) by $-10°$ and scaling factor 0.8. (d) Ground truth MR image overlayed on the fixed image. (e) to (j) Registration results: ELASTIX, AIRNet, SIFT, MIND, CNNFR and TrCNNFR overlayed on the fixed CT.



(a) Clipped fixed CT (b) True clipped MR (c) Initial moving image (d) Ground truth overlay (e) ELASTIX

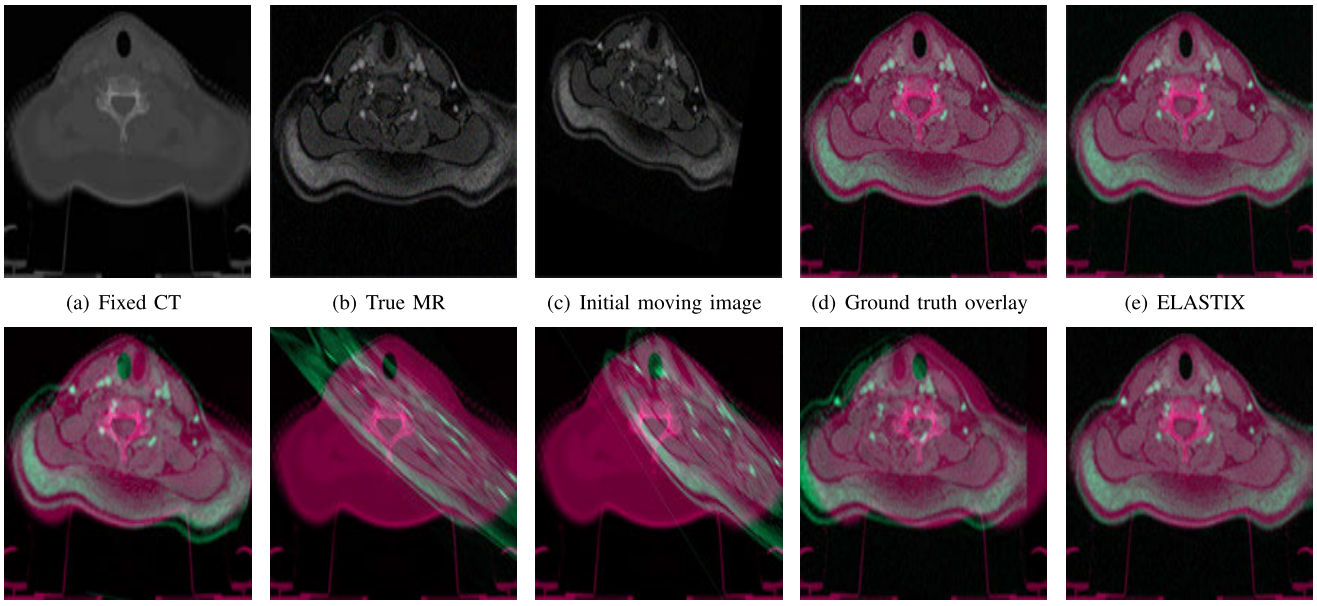
(f) AIRNet (g) SIFT (h) MIND (i) CNNFR (j) TrCNNFR

FIGURE 8. Illustration example of different registration methods for small overlap of the fixed and moving images. (a) Fixed CT with right part clipped. (b) Corresponding MR slice (ground truth) with left part clipped. (c) Initial moving image generated through rotating image (b) by $-10°$ and scaling factor 0.8. (d) Ground truth MR image overlayed on the fixed image. (e) to (j) Registration results: ELASTIX, AIRNet, SIFT, MIND, CNNFR and TrCNNFR overlayed on the fixed CT.

Table 5 and 6 show that in terms of TREs, our methods perform robustly in the cases of unseen body parts.

Fig. 9 shows an example of how well different registration algorithms perform in the application of registering a pair of CT-MR slices from the unseen body parts. In this example, the TREs of the six methods are 0.85, 3.87, 11.40, 9.83, 3.40 and 0.61, respectively. We can see from Fig. 9, the visualization results of the six methods are consistent with their TREs, and the visualized registration accuracy results of TrCNNFR and ELASTIX are comparable.

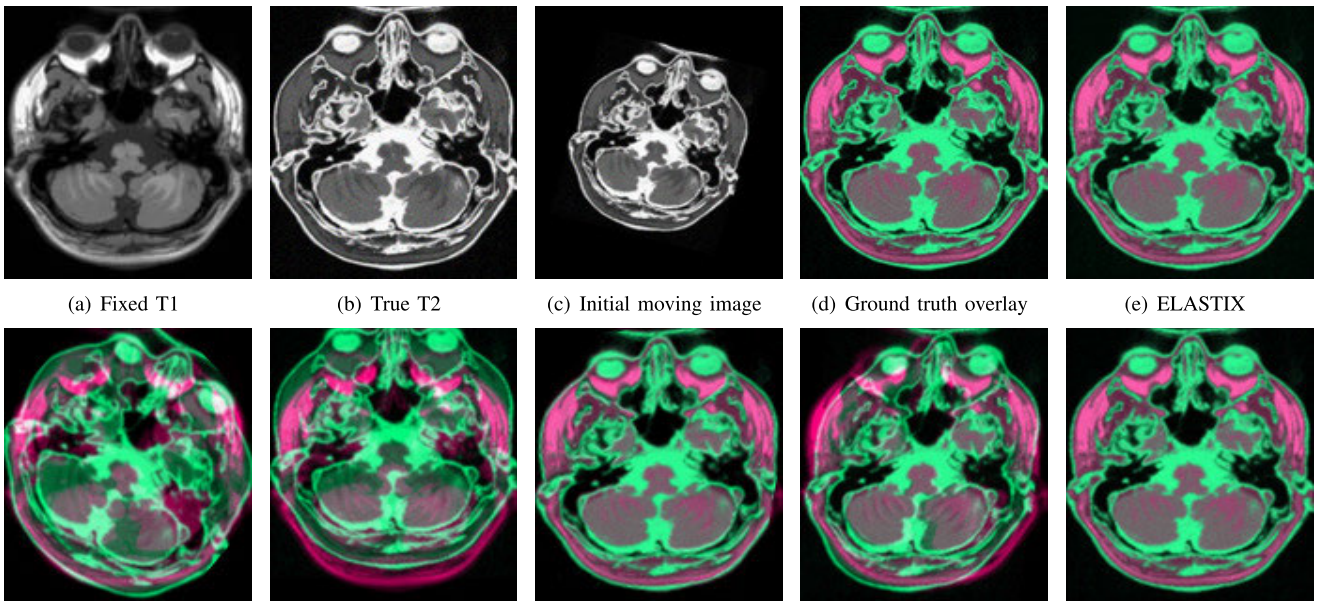
For the second registration task, we used 180 pairs of T1-T2 images from BrainWeb [36]. The quantitative



(a) Fixed CT (b) True MR (c) Initial moving image (d) Ground truth overlay (e) ELASTIX

(f) AIRNet (g) SIFT (h) MIND (i) CNNFR (j) TrCNNFR

FIGURE 9. Illustration example of unseen body parts of different registration methods. (a) Fixed CT. (b) Corresponding MR slice (ground truth). (c) Moving MR generated through rotating image (b) by 15° and scaling factor 0.8. (d) Ground truth MR image overlayed on the fixed image. (e) to (j) Registration results: ELASTIX, AIRNet, SIFT, MIND, CNNFR and TrCNNFR overlayed on the fixed CT.



(a) Fixed T1 (b) True T2 (c) Initial moving image (d) Ground truth overlay (e) ELASTIX

(f) AIRNet (g) SIFT (h) MIND (i) CNNFR (j) TrCNNFR

FIGURE 10. Illustration example of T1-T2 images from BrainWeb of different registration methods. (a) Fixed T1 MR. (b) Corresponding T2 MR slice (ground truth). (c) Moving T2 MR generated through rotating image (b) by 15° and scaling factor 0.8. (d) Ground truth T2 MR image overlayed on the fixed T1 image. (e) to (j) Registration results: ELASTIX, AIRNet, SIFT, MIND, CNNFR and TrCNNFR overlayed on the fixed T1 image.

evaluation results of the six methods are given in Tables 7, 8. In terms of reported TREs in the tables, SIFT and ELASTIX perform better than in table 2, 3, 5 and 6 because of good image quality in BrainWeb dataset. It is noted that our methods obtain comparable registration accuracy than ELASTIX for registering images from unseen image modalities. Fig. 10 illustrates how well different registration algorithms perform in the application of registering a pair of images from the unseen image modalities. In this example, the TREs of the six methods are 0.33, 3.72, 4.88, 0.47, 3.03 and 0.61, respectively.

IV. DISCUSSION

In this work, we proposed contrastive metric learning-based rigid multi-modal medical image registration methods CNNFR and TrCNNFR distilled knowledge from

TABLE 5. TREs of the rotated registration task on image pairs from the unseen body parts.

Rotation Degree	0	5	10	15	20
ELASTIX	0.97 ± 0.43	0.90 ± 0.45	0.96 ± 0.55	1.17 ± 0.57	1.59 ± 0.73
AIRNet	3.17 ± 1.43	3.51 ± 1.45	4.16 ± 1.55	4.57 ± 1.76	4.79 ± 1.83
SIFT	6.78 ± 3.95	7.30 ± 4.18	8.03 ± 4.51	9.69 ± 6.53	11.11 ± 6.95
MIND	5.97 ± 3.43	6.50 ± 2.85	6.96 ± 3.55	7.17 ± 3.77	8.59 ± 4.23
CNNFR	2.37 ± 1.04	2.94 ± 1.13	3.32 ± 1.47	3.86 ± 1.73	4.09 ± 1.58
TrCNNFR	1.16 ± 0.78	1.30 ± 0.90	1.68 ± 1.21	2.33 ± 1.30	2.95 ± 1.55

TABLE 6. TREs of the scaled registration task on image pairs from the unseen body parts.

Scaling Factor	0.7	0.8	0.9	1.0	1.1	1.2
ELASTIX	1.52 ± 0.49	1.12 ± 0.38	0.99 ± 0.36	0.97 ± 0.43	1.20 ± 0.51	1.44 ± 0.58
AIRNet	4.52 ± 1.49	3.12 ± 1.38	2.97 ± 0.96	2.67 ± 0.94	3.15 ± 1.21	3.94 ± 1.38
SIFT	9.55 ± 4.92	10.86 ± 5.46	8.24 ± 4.89	6.78 ± 3.95	8.30 ± 5.36	9.27 ± 4.63
MIND	7.52 ± 2.49	6.82 ± 2.38	5.99 ± 1.86	4.97 ± 1.73	5.60 ± 2.51	6.84 ± 2.98
CNNFR	3.74 ± 1.29	3.04 ± 1.16	2.88 ± 1.37	2.37 ± 1.04	2.86 ± 1.06	3.52 ± 1.15
TrCNNFR	1.87 ± 1.12	1.55 ± 1.03	1.44 ± 0.98	1.16 ± 0.78	1.28 ± 0.50	1.34 ± 1.09

TABLE 7. TREs of the rotated registration task on the T1-T2 images from BrainWeb.

Rotation Degree	0	5	10	15	20
ELASTIX	0.29 ± 0.12	0.35 ± 0.13	0.24 ± 0.11	0.34 ± 0.21	0.42 ± 0.22
AIRNet	1.91 ± 0.62	2.25 ± 0.73	2.84 ± 0.81	3.04 ± 1.21	3.92 ± 1.22
SIFT	2.54 ± 0.68	2.79 ± 0.87	3.28 ± 0.85	4.28 ± 1.50	5.92 ± 1.67
MIND	0.15 ± 0.11	0.36 ± 0.16	0.51 ± 0.32	0.7 ± 0.66	1.95 ± 1.03
CNNFR	1.35 ± 0.93	1.25 ± 0.80	1.99 ± 0.91	2.36 ± 1.23	3.45 ± 1.56
TrCNNFR	0.17 ± 0.15	0.36 ± 0.42	0.71 ± 1.04	1.80 ± 0.77	2.30 ± 1.22

TABLE 8. TREs of the scaled registration task on the T1-T2 images from BrainWeb.

Scaling Factor	0.7	0.8	0.9	1.0	1.1	1.2
ELASTIX	0.63 ± 0.54	0.43 ± 0.01	0.29 ± 0.11	0.29 ± 0.12	0.25 ± 0.11	0.34 ± 0.13
AIRNet	2.95 ± 1.54	2.43 ± 1.01	1.89 ± 0.91	1.69 ± 0.82	1.45 ± 0.70	2.54 ± 1.13
SIFT	4.55 ± 0.71	3.61 ± 0.90	2.62 ± 0.84	2.54 ± 0.68	2.90 ± 1.05	3.71 ± 1.80
MIND	0.82 ± 0.43	0.40 ± 0.21	0.23 ± 0.12	0.15 ± 0.11	0.21 ± 0.12	0.20 ± 0.15
CNNFR	1.58 ± 1.08	1.21 ± 0.86	1.52 ± 0.89	1.35 ± 0.93	1.03 ± 0.90	1.91 ± 1.75
TrCNNFR	0.80 ± 0.62	0.90 ± 0.21	0.38 ± 0.12	0.17 ± 0.15	0.22 ± 0.17	0.31 ± 0.20

natural images. Experiment results demonstrate our proposed method performed robust, accurate compared to the state of the art methods, such as Elastix and AIRNet.

Fig. 4 shows that TrCNNFR and CNNFR perform better than SIFT and MIND regarding key point matching. The learned descriptor proposed in this work using contrastive loss is robust in the task of multi-modal image patch matching. Besides, transfer learning in TrCNNFR from nature images enriches low-level feature variations, and thus TrCNNFR performs more robust than CNNFR training from scratch.

High recall and precision of the matching algorithm play a crucial role in our registration algorithm. Experiment results of image registration in Table 2, 3 and Fig. 6 demonstrate robust matching algorithm performs better in the registration task. Our proposed methods make the matching-based multi-modal image registration feasible comparing to SIFT descriptor and MIND descriptor. Also, our methods outperform intensity-based method (ELASTIX) and supervised deep learning registration method (AIRNet).

Table 4 shows that our proposed method performs about 29 times faster than ELASTIX-based method. At the same time, due to AIRNet is an end-to-end deep learning registration method, it takes the least average running time.

Fig. 7 and Fig. 8 demonstrate the robustness of our proposed algorithm when the fixed and moving images have a small overlap ratio. ELASTIX fails to align two images in the cases of small overlap ratio. ELASTIX utilizes mutual information [37], [38] as cost function under an assumption of global statistical similarity between two images. On the contrary, once we find a sufficient number of corresponding points using our methods, we could robustly register two images. Thus, our method is successful to register two images with small overlap ratio.

From Table 5, 6, 7, 8 and Fig. 9, 10, we observe that our algorithm is capable of generalizing different tasks without retraining the model. The model learns a generic local descriptor which is applicable in other CT-MR applications. From Table 5, 6, 7, 8, it is noted that TrCNNFR performs better than CNNFR. Regarding the SIFT and ELASTIX, we note consistent results in both datasets as in the testing NPC. This is because ELASTIX and SIFT are not able to directly optimize matching features from different modalities. We observe that in the unseen image modalities (referring to Table 7, 8), MIND descriptor-based method attained comparable accuracy to ELASTIX and TrCNNFR. That's because the T1-T2 images from Brainweb are simulated data, which preserves the particular intensity distribution of brain anatomical structure correspondence perfectly. We also observe that in the unseen body parts test and unseen image modalities (referring to Table 5, 6, 7 and 8), learning-based methods perform slightly worse than seen body parts, especially for the AIRNet. Enlarging training dataset to more body parts and image modalities may solve this problem. It is noted that the learning algorithms are a local descriptor learner and thus we only need annotate sparse representative points across different body parts and different image modalities. We plan to investigate this extension in future work.

V. CONCLUSION

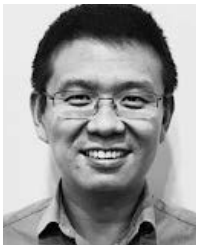
In this work, we presented a matching-based rigid multi-modal medical image registration method. In this framework, the critical component is learning keypoint descriptors using contrastive loss. The experimental results demonstrate that the learning-based keypoint descriptors perform better than handcrafted descriptors in multi-modal registration task. Also, our method outperforms conventional intensity-based image registration method like ELASTIX and deep learning-based registration method like AIRNet regarding registration accuracy, robustness. In the future, we would like to extend our algorithm as followings: (1) We will investigate keypoint detection algorithm. (2) We will investigate end-to-end deep learning keypoint feature-based rigid registration framework

to reduce the running time. (3) We will investigate more image modalities.

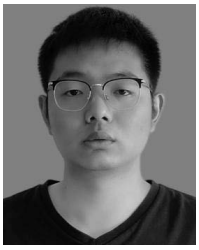
REFERENCES

[1] F. Maes, A. Collignon, D. Vandermeulen, G. Marchal, and P. Suetens, "Multimodality image registration by maximization of mutual information," *IEEE Trans. Med. Imag.*, vol. 16, no. 2, pp. 187–198, Apr. 1997.
[2] A. Roche, G. Malandain, X. Pennec, and N. Ayache, "The correlation ratio as a new similarity measure for multimodal image registration," in *Medical Image Computing and Computer-Assisted Intervention—MICCAI*. Berlin, Germany: Springer-Verlag, 1998, pp. 1115–1124.
[3] M. Toews, L. Zöllei, and W. M. Wells, "Feature-based alignment of volumetric multi-modal images," in *Proc. Int. Conf. Inf. Process. Med. Imag.*, 2013, pp. 25–36.
[4] K. Padgett, R. Stoyanova, P. Johnson, J. Piper, A. Javorek, N. Dogan, and A. Pollack, "SU-F-BRF-10: Deformable MRI to CT validation employing same day planning MRI for surrogate analysis," *Med. Phys.*, vol. 41, p. 401, Jun. 2014.
[5] J. Woo, M. Stone, and J. L. Prince, "Multimodal registration via mutual information incorporating geometric and spatial context," *IEEE Trans. Image Process.*, vol. 24, no. 2, pp. 757–769, Feb. 2015.
[6] M. Simonovsky, B. Gutiérrez-Becker, D. Mateus, N. Navab, and N. Komodakis, "A deep metric for multimodal registration," in *Medical Image Computing and Computer-Assisted Intervention—MICCAI*. Cham, Switzerland: Springer, 2016.
[7] S. Miao, Z. J. Wang, and R. Liao, "A CNN regression approach for real-time 2D/3D registration," *IEEE Trans. Med. Imag.*, vol. 35, no. 5, pp. 1352–1363, May 2016.
[8] B. D. de Vos, F. F. Berendsen, M. A. Viergever, M. Staring, and I. Išgum, "End-to-end unsupervised deformable image registration with a convolutional neural network," in *Proc. Deep Learn. Med. Image Anal. Multimodal Learn. Clin. Decis. Support*. Cham, Switzerland: Springer, 2017, pp. 204–212.
[9] X. Cao, J. Yang, J. Zhang, D. Nie, M. Kim, Q. Wang, and D. Shen, "Deformable image registration based on similarity-steered CNN regression," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2017, pp. 300–308.
[10] X. Yang, R. Kwitt, M. Styner, and M. Niethammer, "Quicksilver: Fast predictive image registration—A deep learning approach," *NeuroImage*, vol. 158, pp. 378–396, Sep. 2017.
[11] M.-M. Rohé, M. Datar, T. Heimann, M. Sermesant, and X. Pennec, "SVF-Net: Learning deformable image registration using shape matching," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2017, pp. 266–274.
[12] J. Zheng, S. Miao, W. Z. Jane, and R. Liao, "Pairwise domain adaptation module for CNN-based 2-D/3-D registration," *J. Med. Imag.*, vol. 5, no. 2, 2018, Art. no. 021204.
[13] K. R. Padgett, R. Stoyanova, S. Pirozzi, P. Johnson, J. Piper, N. Dogan, and A. Pollack, "Validation of a deformable MRI to CT registration algorithm employing same day planning MRI for surrogate analysis," *J. Appl. Clin. Med. Phys.*, vol. 19, no. 2, pp. 258–264, 2018.
[14] Y. Hu, M. Modat, E. Gibson, W. Li, N. Ghavami, E. Bonmati, G. Wang, S. Bandula, C. M. Moore, S. Ourselin, J. A. Noble, D. C. Barratt, T. Vercauteren, and M. Emberton, "Weakly-supervised convolutional neural networks for multimodal image registration," *Med. Image Anal.*, vol. 49, pp. 1–13, Oct. 2018.
[15] U. Kruger, G. Haskins, and P. Yan, "Deep learning in medical image registration: A survey," 2019, *arXiv:1903.02026*. [Online]. Available: https://arxiv.org/abs/1903.02026
[16] J. P. W. Pluim, J. B. A. Maintz, and M. A. Viergever, "Mutual-information-based registration of medical images: A survey," *IEEE Trans. Med. Imag.*, vol. 22, no. 8, pp. 986–1004, Aug. 2003.
[17] S. Klein, M. Staring, K. Murphy, M. A. Viergever, and J. P. Pluim, "Elastix: A toolbox for intensity-based medical image registration," *IEEE Trans. Med. Imag.*, vol. 29, no. 1, pp. 196–205, Jan. 2010.
[18] D. Glodeck, J. Hesser, and L. Zheng, "Potential of metric homotopy between intensity and geometry information for multi-modal 3D registration," *Zeitschrift Medizinische Phys.*, vol. 28, pp. 325–334, Feb. 2018.
[19] J. P. W. Pluim, J. B. A. Maintz, and M. A. Viergever, "Image registration by maximization of combined mutual information and gradient information," *IEEE Trans. Med. Imag.*, vol. 19, no. 8, pp. 809–814, Aug. 2000.
[20] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.
[21] Q. Li, G. Wang, J. Liu, and S. Chen, "Robust scale-invariant feature matching for remote sensing image registration," *IEEE Geosci. Remote Sens. Lett.*, vol. 6, no. 2, pp. 287–291, Apr. 2009.
[22] D. Zhao, Y. Yang, Z. Ji, and X. Hu, "Rapid multimodality registration based on MM-SURF," *Neurocomputing*, vol. 131, pp. 87–97, May 2014.
[23] S. W. Teng, M. T. Hossain, and G. Lu, "Multimodal image registration technique based on improved local feature descriptors," *J. Electron. Imag.*, vol. 24, no. 1, 2015, Art. no. 013013.
[24] G. Lv, S. W. Teng, and G. Lu, "Enhancing SIFT-based image registration performance by building and selecting highly discriminating descriptors," *Pattern Recognit. Lett.*, vol. 84, pp. 156–162, Dec. 2016.
[25] W. Ma, Z. Wen, Y. Wu, L. Jiao, M. Gong, Y. Zheng, and L. Liu, "Remote sensing image registration with modified sift and enhanced feature matching," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 1, pp. 3–7, Jan. 2017.
[26] Z. Hossein-Nejad and M. Nasri, "An adaptive image registration method based on SIFT features and RANSAC transform," *Comput. Elect. Eng.*, vol. 62, pp. 524–537, Aug. 2017.
[27] C. Zhao, H. Zhao, J. Lv, S. Sun, and B. Li, "Multimodal image matching based on multimodality robust line segment descriptor," *Neurocomputing*, vol. 177, pp. 290–303, Feb. 2016.
[28] F. Ye, Y. Su, H. Xiao, X. Zhao, and W. Min, "Remote sensing image registration using convolutional neural network features," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 2, pp. 232–236, Feb. 2018.
[29] M. P. Heinrich, M. Jenkinson, M. Bhushan, T. Matin, F. V. Gleeson, S. M. Brady, and J. A. Schnabel, "MIND: Modality independent neighbourhood descriptor for multi-modal deformable registration," *Med. Image Anal.*, vol. 16, no. 7, pp. 1423–1435, 2012.
[30] S. Chopra, R. Hadsell, and Y. LeCun, "Learning a similarity metric discriminatively, with application to face verification," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 1, Jun. 2005, pp. 539–546.
[31] M. Brown, G. Hua, and S. Winder, "Discriminative learning of local image descriptors," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 1, pp. 43–57, Jan. 2011.
[32] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[33] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. B. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," 2014, *arXiv:1408.5093*. [Online]. Available: https://arxiv.org/abs/1408.5093
[34] M. A. Fischler and R. C. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," in *Readings in Computer Vision*. San Francisco, CA, USA: Morgan Kaufmann, 1987, pp. 726–740.
[35] E. Chee and Z. Wu, "AIRNet: Self-supervised affine registration for 3d medical images using neural networks," 2018, *arXiv:1810.02583*. [Online]. Available: https://arxiv.org/abs/1810.02583
[36] D. L. Collins, A. P. Zijdenbos, V. Kollokian, J. G. Sled, N. J. Kabani, C. J. Holmes, and A. C. Evans, "Design and construction of a realistic digital brain phantom," *IEEE Trans. Med. Imag.*, vol. 17, no. 3, pp. 463–468, Jun. 1998.
[37] P. Viola and W. M. Wells, III, "Alignment by maximization of mutual information," *Int. J. Comput. Vis.*, vol. 24, no. 2, pp. 137–154, Sep. 1997.
[38] C. E. Shannon, W. Weaver, and N. Wiener, "The mathematical theory of communication," *Phys. Today*, vol. 3, no. 9, pp. 31–32, 1950.

JINRONG HU received the B.Sc. and M.Sc. degrees from Sichuan Normal University Chengdu, Sichuan, China, in 2005 and 2008 respectively, and the Ph.D. degree from Sichuan University, Chengdu, in 2012, all in computer science. She is currently an Associate Professor with the Chengdu University of Information Technology and Xihua University. Her current research interests include the image processing, artificial intelligence, and medical image analysis.



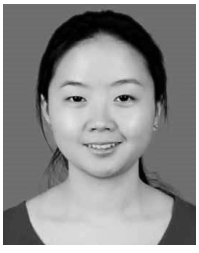
SHANHUI SUN received the Ph.D. degree from the Department of Electrical and Computer Engineering (ECE), The University of Iowa, in 2012, and the M.S. degree in electrical and automation engineering from Tianjin University, in 2006. He was a Staff Scientist with Siemens Corporate Research, Princeton. He is currently a Principal Scientist with CuraCloud Corporation.



XIAODONG YANG received the B.S. degree in computer science and technology from the Chengdu University of Information Technology, Chengdu, China, in 2017, where he is currently pursuing the M.S. degree in computer technology. His current research interests include computer vision, image processing, and deep learning.

SHUANG ZHOU received the bachelor's degree in software engineering from the University of Electronic Science and Technology of China (UESTC), in 2009, and the master's degree in artificial intelligence and the Ph.D. degree from the Department of Data Science and Knowledge Engineering (DKE), Maastricht University, The Netherlands, in 2012 and 2017, respectively. She is currently a Data Scientist with Philips Research China. Her current research interests include transfer learning, conformal prediction, and their applications.

XIN WANG received the Ph.D. degree in computer science from the University at Albany, State University of New York, in 2015. He is currently a Senior Machine Learning Scientist with CuraCloud Corporation. His current research interests include artificial intelligence, machine learning, and computer vision.



YING FU received the Ph.D. degree from Sichuan University, in 2014. She is currently an Associate Professor with the Chengdu University of Information Technology. Her current research interest includes inverse problem of image processing, specially, the application of nonparametric Bayesian method in image restoration.

JILIU ZHOU received the B.Sc. degree in electronic and computer science from Sichuan University, in 1985, the M.Sc. degree in electronic and computer science from Tsinghua University, in 1988, and the Ph.D. degree from Sichuan University, in 1999. He was a Full Professor with Sichuan University, in 1999. He is currently with Sichuan University and the Chengdu University of Information Technology (CUIT), as a Full Professor. He is currently the Director of the Collaborative Innovation Center for Image and Geospatial Information (CICIGI). He has published more than 200 journal articles.

YOUBING YIN received the Ph.D. degree from The University of Iowa. He was the Senior Research Scientist with VIDA Diagnostics. He is currently the Vice President of Research and Development, CuraCloud Corporation. His research interest includes the development and commercialization of machine learning and imaging-based solutions in healthcare. He is familiar with strict developing procedures in this highly regulated industry (FDA, CE, and CFDA). He was a Key Developer of a FDA approved quantitative lung imaging software, which received the Best Product Award from the European Respiratory Society.

KUNLIN CAO received the Ph.D. degree in electrical and computer engineering from The University of Iowa, USA. She was a Lead Scientist with the Biomedical Image Analysis Laboratory, GE Global Research Center. She has over 10 years' of research experience in medical image analysis, and extensively involved in image segmentation, image registration, feature extraction, functional information quantification, motion detection, and object localization/tracking for real-time image guided applications. She published over 60 articles at international journals and conferences in medical image analysis area, which were cited more than 1000 times. Her current research interests include multi-dimensional image and signal processing, medical image analysis (CT, Ultrasound, MR, and PET), artificial intelligence, and computer vision. She was with CuraCloud Corporation, involved in a variety of research projects and product development in the field of quantitative assessment of anatomical and functional changes toward precise disease diagnosis/prognosis/treatment planning using artificial intelligence technologies.

QI SONG received the bachelor's degree from the University of Electronic Science and Technology, in 2003, the master's degree from Tsinghua University, in 2006, and the Ph.D. degree from The University of Iowa, USA, in 2011. He was a Research Assistant with The University of Iowa, in 2011. He became a Scientist with the General Electric Company's Global Research and Development Center, New York, in 2014. He was a Senior Scientist with HeartFlow Corporation, USA, from 2014 to 2015, and a Founder and the CEO of CuraCloud Corporation, Seattle, USA, from 2016 to 2017. Since 2017, he has been the Founder and the General Manager of ShenzhenKeya Medical Technology Company Ltd.

XI WU received the B.S. degree in communication engineering from Sichuan University, Chengdu, China, in 2003, the M.S. degree in communication engineering from the University of Electronic Science and Technology of China, Chengdu, in 2006, and the Ph.D. degree in computer science from Sichuan University, in 2012. He is currently a Professor with the Department of Computer Science, Chengdu University of Information Technology, Chengdu. His current research interests include image processing and computer vision.