BioMedical Engineering
OnLine

RESEARCH **Open Access**

macJNet: weakly-supervised multimodal image deformable registration using joint learning framework and multi-sampling cascaded MIND

Zhiyong Zhou[1,3†], Ben Hong[2†], Xusheng Qian[1,3], Jisu Hu[1,3], Minglei Shen[2], Jiansong Ji[4*] and Yakang Dai[1,3*]

†Zhiyong Zhou and Ben Hong are co-first authors.

*Correspondence:
jjstcty@sina.com; daiyk@sibet.ac.cn

[1] Suzhou Institute of Biomedical Engineering and Technology, Chinese Academy of Sciences, Suzhou, Jiangsu, China
[2] School of Electronic and Optical Engineering, NanJing University of Science and Technology, Nanjing, Jiangsu, China
[3] School of Biomedical Engineering (Suzhou), Division of Life Sciences and Medicine, University of Science and Technology of China, Suzhou, Jiangsu, China
[4] Key Laboratory of Imaging Diagnosis and Minimally Invasive Intervention Research, The Fifth Affiliated Hospital of Wenzhou Medical University, Lishui, Zhejiang, China

Abstract

Deformable multimodal image registration plays a key role in medical image analysis. It remains a challenge to find accurate dense correspondences between multimodal images due to the significant intensity distortion and the large deformation. macJNet is proposed to align the multimodal medical images, which is a weakly-supervised multimodal image deformable registration method using a joint learning framework and multi-sampling cascaded modality independent neighborhood descriptor (macMIND). The joint learning framework consists of a multimodal image registration network and two segmentation networks. The proposed macMIND is a modality-independent image structure descriptor to provide dense correspondence for registration, which incorporates multi-orientation and multi-scale sampling patterns to build self-similarity context. It greatly enhances the representation ability of cross-modal features in the registration network. The semi-supervised segmentation networks generate anatomical labels to provide semantics correspondence for registration, and the registration network helps to improve the performance of multimodal image segmentation by providing the consistency of anatomical labels. 3D CT-MR liver image dataset with 118 samples is built for evaluation, and comprehensive experiments have been conducted to demonstrate that macJNet achieves superior performance over state-of-the-art multi-modality medical image registration methods.

Keywords: Deformable registration, Multimodal, Image descriptor, Joint learning, Semi-supervised segmentation

Introduction

Multimodal medical image registration aims to establish anatomical correspondences between multimodal images, which plays an important role in assisted diagnosis, image-guided ablation, and surgical navigation. Medical image registration is a high-dimensional optimization task to estimate the dense deformation fields. With the recent advances in data driven learning, deep learning-based registration methods have achieved comparable accuracy with a significantly higher inference speed. In general, deep learning-based registration could be categorized into fully-supervised registration, unsupervised registration and weak-supervised registration from the perspective of the utilization of the ground-truth.

Fully-supervised registration

Inspired by the FlowNet for vector flow estimation [1], fully-supervised image registration methods [2–4] consider image registration as a regression problem to predict deformation fields for matching the ground-truth. Fully-supervised registration imports image pairs and dense correspondence to learn the spatial mapping between images, and directly predicts deformation fields in the inference stage. It makes the fully-supervised registration a modality-independent registration method. However, it is challenging to find the accurate dense correspondence between medical images. Fan et al. [5] proposed brain image registration networks (BIRNET) to guide the training process in fully-supervised learning using a dual supervision loss to measure the difference between the generated deformation field and the real deformation field. Cao et al. [6] cascaded Syn [7] and Demons [8] to obtain the deformations used as ground-truth for CNN training. Some methods generate the artificially synthesized images to simulate the deformation fields [9], which solves the problem of getting dense correspondence between images. However, the authenticity problem of synthesized warped images would degrade registration performance.

Unsupervised registration

Unsupervised registration methods do not require ground-truth deformation fields [10, 11], which consider image registration as a loss function minimization problem and use a differentiable warping module with the spatial transformer network (STN) [11] to warp the moving image in the training procedure. The image similarity metric and regularization are usually incorporated into the loss function to optimize the registration network. Learning the cross-modality representation through network training or designing elaborated modality-independent similarity metrics are two alternative ways for multimodal registration.

In the first way, Balakrishnan et al. [11] proposed the first unsupervised learning registration method (VoxelMorph) for mono-modality registration. Mok and Chung [12] further improved its performance by adding the symmetric diffeomorphic properties into the network. To efficiently train a medical image registration network, DeepFLASH [13] computes the deformation fields via utilizing low-dimensional band-limited space. Yan et al. [14] first proposed the adversarial image registration framework, which performs image registration tasks through a generator and evaluates the quality of the warped images by a discriminator. Kim et al. [15] proposed a fully convolutional self-similarity to find dense semantic correspondence in mono-modality registration. A recent trend for multimodal image registration takes advantages of image to image translation [16], generative adversarial networks (GANs) convert the multimodal registration into a simpler unimodal task by learning transferable representations from multimodal images. Fan et al. [17] further extended this work to both unimodal and multimodal registration.

However, image translation is a challenging topic by itself, the main challenges for GANs-based registration include: it may inevitably produce artificial features [18] and achieving Nash equilibrium in training procedure [19].

In other way, some methods attempt to elaborately design cross-modal descriptors as a similarity metric to represent the modality-independent structure features for multimodal registration. Schechtman and Irani [20] introduced the local self-similarity (LSS) descriptor for multimodal image matching address the problem of multimodal appearance and shape change. Heinrich et al. [21] proposed a modality-independent neighborhood descriptor (MIND) based on self-similarity theory [20], which calculates the difference between patches within a local neighborhood. Some other LSS-based methods are also introduced to represent the cross-modal dense correspondence [22, 23]. Kim et al. proposed deep self-correlation (DSC) [24] to estimate cross-modal dense correspondences inspired by LSS and DSC has demonstrated its high accuracy on aligning multimodal image. Fully convolutional self-similarity (FCSS) [15] formulates LSS within a fully convolutional network to simultaneously learn the patch sampling patterns and self-similarity measures. Although FCSS dramatically improved performance for object-level semantic correspondence, it cannot deal with complex geometric variations, which frequently appears in medical image registration.

Weakly-supervised registration

Weakly-supervised registration usually uses anatomical segmentation labels as semantic prior information to improve the registration performance. However, manual delineation of anatomical labels is a time-consuming and laborious work. To address the problem of insufficient labels, the joint learning framework for registration and segmentation has been proposed [25–27], in which the registration and segmentation network are alternately optimized during the training procedure. Some label-driven weakly-supervised methods have also been proposed [28, 29] by exploiting the auxiliary anatomical information and the invertible transformation. In the joint learning framework, the anatomy labels created by the segmentation network provide semantic prior knowledge to guide dense correspondence mapping for the registration network [25]. The registration provides the consistency of segmentation labels by mapping the warped image to the fixed image, which is an effective way to improve the segmentation performance of multimodal images. The registration and segmentation networks are iteratively optimized in an end-to-end manner to simultaneously improve the performances of registration and segmentation [30]. However, the joint learning framework still confronts the following problems. For registration network, it is a challenge on how to utilize semantic labels to provide sufficient dense correspondence between multimodal images [31], which leads to the low quality of registration in interior of large tissues, such as liver. For segmentation network, it is a challenge to generate the consistent labels for multimodal images with few manual labels.

In general, the existing registration methods cannot accurately align the multimodal images since they cannot learn the cross-modality dense correspondence to handle complex and large deformation. In this paper, macJNet is proposed as a novel multimodality registration method, which is weakly-supervised multimodal image deformable registration using joint learning framework and multi-sampling cascaded modality independent neighborhood descriptor (macMIND). The key idea behind macJNet is to learn (or extract) different levels of prior knowledge to guide the registration: anatomical labels are predicted by segmentation networks as semantic information to provide global sparse correspondences for registration, and the macMIND is extracted as context information to provide local dense correspondences for registration. Our contributions are summarized as follows.

(1) A novel weakly-supervised multimodal image deformable registration methodology using a joint learning framework (macJNet) is proposed for multimodal registration. The macJNet consists of a registration network and two segmentation networks, which are iteratively optimized in a single end-to-end framework. Segmentation networks provide semantic anatomical labels for weakly-supervised registration by few-label learning; registration network improves the performance of segmentation results by enforcing cross-modality consistency based on deformable spatial mapping.

(2) Multi-sampling cascaded modality independent neighborhood descriptor (macMIND) is proposed to establish dense correspondences between multimodal images for registration. macMIND builds the local self-similarity context by multi-orientation and multi-scale sampling in a supporting window, which enriches the modality-independence contextual information to characterize cross-modality anatomical structures. An efficient computational scheme for macMIND in a convolutional manner is also proposed.

(3) Dual similarity-based loss function is introduced to optimize macJNet. The dual similarity incorporates macMIND and DSC, in which macMIND represents the similarity of modality-independent context to find dense correspondence and DSC represents the similarity of semantic labels structures to find sparse correspondence of tissue boundaries.

The paper is organized as follows. “Experiments” presents the proposed methodology and its implementation. The experiments results are given in “Methodology”. Conclusion and discussion are given in "Methodology".

Experiments

Medical image data and evaluation metrics

118 pairs of CT-MR liver images are used to evaluate the proposed method. All images are collected from Lishui Central Hospital. The characteristic of dataset is listed in Table 1. All anatomy labels (liver labels and tumor labels) and anatomical landmarks

Table 1 The characteristic of the dataset

Symbol	MR image	CT image
Modality	T1(no contrast)	no contrast
FOV	288 × 288	512 × 512
Resolution (mm^3)	1.146 × 1.146 × 3	0.664 × 0.664 × 5
Scanner	Siemens	Philips

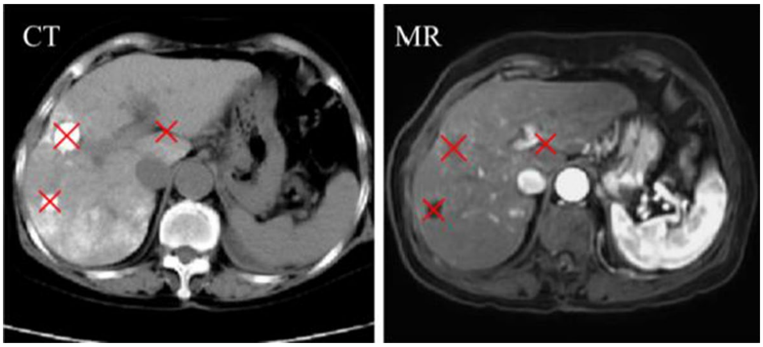


Fig. 1 The identified landmarks (the central location of tumor and hepatic fissures) in CT and MR images. The TRE is defined as Euclidean distance between the corresponding landmarks

(Fig. 1) are executed by radiologists. 90 pairs are selected randomly and assigned into training cohort, and the remaining 28 pairs are assigned into the testing cohort. macJNet is optimized by five-fold cross-validation on the training cohort.

To quantitatively verify the effectiveness of macJNet, target registration error (TRE), Dice similarity coefficient (DSC), 95% Hausdorff distance (Hd_{95}), mutual information (MI), and structural similarity (SSIM) are used to evaluate the registration accuracy. TRE, DSC and Hd_{95} are used to evaluate the accuracy of tumor and liver registration; MI and SSIM are used to evaluate the registration quality over the entire image domain.

Mutual information is a common similarity metric for multimodal image registration, which indicates the similarity of two images. The mutual information is defined as:

$$\mathrm{MI}(I_F, I_M) = \sum_{I_F, I_M} p(I_F, I_M) \log \frac{p(I_F, I_M)}{p(I_F)p(I_M)}, \tag{1}$$

where the probability $p(I)$ is the probability distribution of the voxel values in image I, and the probability $p(I_F, I_M)$ is the joint distribution of the intensities of two images.

SSIM is a metric to measure the structural similarity between two images, which mainly focus on structural information (such as shapes and position). The range of SSIM is from 0 to 1, a higher value implies a higher similarity [32]. SSIM has been applied as similarity metric in a GAN-based brain multimodal registration [33]. SSIM is defined as

$$\mathrm{SSIM} = \frac{\left(2\bar{I}_F\bar{I}_M + c_1\right) + \left(2\sigma_{M-F} + c_2\right)}{\left(\bar{I}_F^2 + \bar{I}_M^2 + c_1\right)\left(\sigma_F^2 + \sigma_M^2 + c_2\right)}, \tag{2}$$

where $\bar{I}$ symbolizes the mean voxel value of the given image; σ is the standard deviation of the image; $\sigma_{M\text{-}F}$ is the covariance of multimodal image pair; c_1 and c_2 are constant values.

Registration results

Implementation

In light of the limited GPU computing resources, the liver images are resampled into $128 \times 128 \times 96$ and then input into macJNet for training and inference. The output deformation fields and warped images would be up-sampled to original size. The

Table 2 Comparisons of registration results (mean ± std)

Methods	Tumor			Liver			Image		Time (s)
	TRE (mm)	DSC (%)	Hd_{95} (mm)	TRE (mm)	DSC (%)	Hd_{95}(mm)	MI (%)	SSIM (%)	
Affined	7.38 ± 3.56	46.81 ± 32.67	7.70 ± 3.27	7.01 ± 2.95	90.94 ± 1.38	6.67 ± 1.51	32.11 ± 6.96	34.29 ± 10.54	22.5 ± 3.8
Elastix	5.23 ± 1.49	53.27 ± 23.47	7.13 ± 2.03	5.45 ± 2.80	93.54 ± 1.12	5.33 ± 1.22	34.67 ± 12.85	38.42 ± 10.32	84.1 ± 7.2
VoxelMorph	5.89 ± 3.17	50.95 ± 28.90	7.06 ± 1.93	6.83 ± 2.85	93.18 ± 1.28	5.79 ± 1.64	35.88 ± 6.74	36.83 ± 11.77	0.20 ± 0.01
LapIRN	5.48 ± 2.24	52.51 ± 22.64	7.03 ± 1.71	5.51 ± 1.39	93.64 ± 1.13	5.52 ± 1.37	42.03 ± 4.71	49.03 ± 12.66	0.17 ± 0.02
macJNet	**5.05 ± 1.77**	**55.20 ± 18.77**	**6.71 ± 1.97**	**4.83 ± 1.49**	**94.75 ± 0.82**	**4.53 ± 1.11**	**44.12 ± 4.63**	**54.43 ± 11.62**	0.18 ± 0.02

Bold values indicate better results than other methods

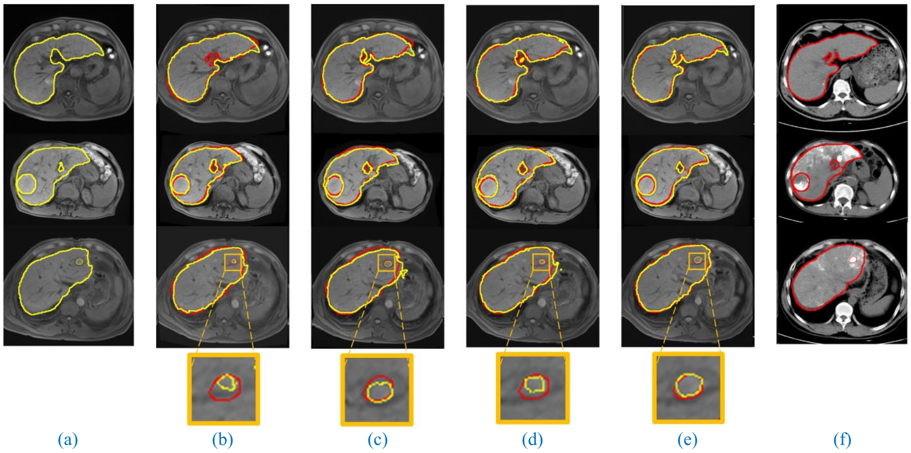


Fig. 2 Visualization of registration on three samples in the test dataset. The left and right columns show the moving and fixed images, respectively. The middle four columns show the results of Elastix, VoxelMorph, LapIRN and macJNet in sequence. **a** MR image (moving image), **b** registration results of Elastix, **c** registration results of VoxelMorph, **d** registration results of LapIRN, **e** registration results of macJNet, **f** CT image (fixed image)

Table 3 Registration with different descriptors in joint learning framework (mean ± std)

Metric	Tumor			Liver			Image	
	TRE (mm)	DSC (%)	Hd_{95} (mm)	TRE (mm)	DSC (%)	Hd_{95}(mm)	MI (%)	SSIM (%)
Affined	7.38 ± 3.56	46.81 ± 32.67	7.70 ± 3.27	7.01 ± 2.95	90.94 ± 1.38	6.67 ± 1.50	32.11 ± 6.96	34.29 ± 10.54
MI	6.03 ± 1.29	50.42 ± 25.68	7.33 ± 2.40	6.17 ± 2.24	92.89 ± 1.95	5.64 ± 1.12	43.65 ± 5.05	42.76 ± 9.78
MIND	5.64 ± 1.51	51.58 ± 23.05	7.18 ± 1.94	5.69 ± 1.41	94.61 ± 1.07	4.77 ± 1.08	42.72 ± 4.85	49.30 ± 12.67
mac-MIND	**5.05 ± 1.79**	**55.20 ± 18.77**	**6.71 ± 1.97**	**4.83 ± 1.49**	**94.75 ± 0.82**	**4.53 ± 1.11**	**44.12 ± 4.63**	**54.43 ± 11.62**

Bold values indicate better results than other methods

Reg-SubNet is pre-trained in an unsupervised manner, and Seg-SubNets is pre-trained in cycle self-training with CT and MR image. 30% liver labels are used to train the Seg-SubNets for guiding registration, and the tumor labels are only used as ground-truth to evaluation the accuracy of registration. The learning rate is set to 2×10^{-5} in registration and 1×10^{-5} in segmentation, batch size is 1, epoch number is 200. The learning rate in registration network is larger than that in segmentation due to the convergence of segmentation is faster than registration. Adam is used as optimizer in these networks. In our experiments, the hyper-parameters are: $K = 2$, $\alpha_1 = 0.3$, $\alpha_2 = 0.7$ in Eq. (11); $\lambda_{\text{sim}} = 20$, $\lambda_{\text{label}} = 2$ and $\lambda_{\text{smo}} = 0.5$ in loss function. $L = 2$-pixel distance, $R_1 = R_2 = 5$-pixel. The joint training cost around 16 h to reach convergence, while it only cost about 0.18 s to complete deformation prediction for an image pair.

To evaluate the registration performance, macJNet is performed to compared with the well-performed methods: Elastix [34], VoxelMorph [11], and LapIRN [35]. Elastix is a classic traditional registration method using mutual information-based multimodal similarity metric, and 3-level pyramid in Elastix is used in the experiments. VoxelMorph is a CNN-based unsupervised registration method, which is aimed to mono-modality image registration. VoxelMorph with MIND-based loss function is applied to multimodal

Table 4 Performance of macMIND in registration network (mean ± std)

Methods	Tumor			Liver			Image	
	TRE (mm)	DSC (%)	Hd_{95} (mm)	TRE (mm)	DSC (%)	Hd_{95}(mm)	MI (%)	SSIM (%)
MIND	5.48 ± 2.24	52.51 ± 22.64	7.03 ± 1.71	5.51 ± 1.39	93.64 ± 1.13	5.52 ± 1.37	42.03 ± 4.71	49.03 ± 12.66
macMIND	**5.19 ± 1.34**	**56.09 ± 19.05**	**6.53 ± 2.18**	**4.90 ± 1.48**	93.64 ± 1.07	**5.39 ± 1.14**	**43.76 ± 4.83**	**53.00 ± 11.64**

Bold values indicate better results than other methods

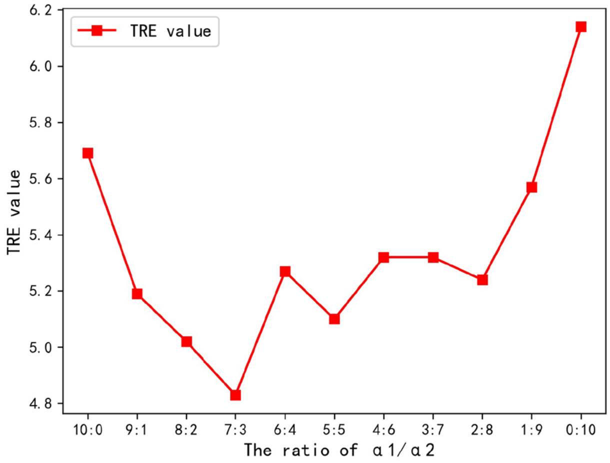


Fig. 3 Optimal ratio of scale weight α_1 and α_2. The horizontal axis indicates the ratio of α_1/α_2, where α_2 is the weight of large sampling window, and α_1 is the weight of small sampling window

registration. The configuration of VoxelMorph is as follows: learning rate of 1×10^{-4}, regularization parameter of 1, batch size of 1, and the number of epochs of 800. LapIRN is a CNN-based registration method, which divides the image into three resolutions and performs registration layer by layer. LapIRN is also applied as a baseline network in RegSubNet. The parameters of configuration are set same as VoxelMorph. All deep learning-based methods are implemented by Pytorch on a single Nvidia Telsa V100 GPU with 16G memory. Elastix registration running on AMD Ryzen 5 4600H CPU. Affine alignment for each image pair is pre-performed using Elastix to reduce the position deviation.

Multimodal image registration results

As shown in Table 2, four deformable registration methods are compared with the metrics of TRE, DSC, Hd_{95}, MI, SSIM and inference time. In terms of tumor registration, it is observed that macJNet achieves better registration performance (TRE = 5.05 mm, DSC = 55.20%, Hd_{95} = 6.71 mm) than Elastix, VoxelMorph and LapIRN. In terms of liver registration, macJNet (TRE = 4.83 mm, DSC = 94.75%, Hd_{95} = 4.53 mm, MI = 44.12%, SSIM = 54.43%) also outperforms other competitive methods in all evaluation metrics. This statistical result demonstrates that macMIND and consistency constraint simultaneously improve the global registration accuracy and local accuracy. Figure 2 intuitively shows the visual comparisons of registration results using different methods, where macJNet optimizes the deformation both in tissue boundary and internal organs. The Elastix outperforms VoxelMorph at tumor alignment with the metric of TRE and DSC and liver alignment with all evaluation metrics. In addition, the inference time of

Table 5 Performance of macJNet and Reg-SubNet with macMIND (mean ± std)

Methods	Tumor			Liver			Image	
	TRE (mm)	DSC (%)	Hd_{95} (mm)	TRE (mm)	DSC (%)	Hd_{95}(mm)	MI (%)	SSIM (%)
Reg-SubNet	5.19 ± 1.34	56.09 ± 19.05	6.53 ± 2.18	4.90 ± 1.48	93.64 ± 1.07	5.39 ± 1.14	43.76 ± 4.83	53.00 ± 11.64
macJNet	**5.05 ± 1.77**	55.20 ± 18.77	6.71 ± 1.96	**4.83 ± 1.49**	**94.75 ± 0.82**	**4.53 ± 1.11**	**44.12 ± 4.63**	**54.43 ± 11.62**

Bold values indicate better results than other methods

Table 6 Performance of macMIND and MIND in deep learning-based registration (mean ± std)

Methods	$\det(J_{\varphi}(p)) < 0$ (‱)
Reg-SubNet with MIND	0.92 ± 0.45
Reg-SubNet with macMIND	**0.63 ± 0.22**
Joint learning framework with MIND	0.96 ± 0.47
Joint learning framework with macMIND (macJNet)	**0.74 ± 0.24**

Bold values indicate better results than other methods

macJNet is comparable to other deep learning-based methods and over 400 times faster than Elastix. The affine registration is listed as a reference to obviously compare the performance of registration methods. It should be noted that clinical medical images are used (slice thickness is larger than 3 mm) in the experiment, which takes an adverse impact on registration result. However, macJNet still accurately matches the multimodal images, and outperforms the competitive methods.

Ablation studies

Evaluation of macMIND To verify the effectiveness of our proposed macMIND in the macJNet, local mutual information (MI) [45], MIND [20] and macMIND are incorporated, respectively, into the macJNet to compare the performance of these modality-independent image descriptors. Table 3 shows the results of the competitive image descriptors for CT-MR deformable registration, which shows the proposed macMIND achieves the best performance in all evaluation metrics for global alignment and local deformation. macMIND have an ability to describes complex cross-modality image structures and their geometrical variants due to its multi-sampling patterns in self-similarity context. Moreover, macMIND also could robustly reflects the large deformation vis multi-scale sampling and cascaded extractions.

Compared with MIND in the joint learning framework, macMIND improves 10.34% for TRE, and 3.62% for DSC, and 6.59% for Hd_{95} in the local (tumor) registration; improves 15.05% for TRE, 0.14% for DSC, 5.03% for Hd_{95}, and 5.13% for SSIM in organ (liver) alignment. The statistical results demonstrate that macMIND is an outstanding descriptor to represent modality-independent image structures.

Furtherly, the effectiveness of macMIND is evaluated in registration network with the unsupervised learning manner. The statistical results of macMIND and MIND are listed in Table 4. It is observed that macMIND significantly improves the performance of registration in almost all evaluation metrics. Specifically, macMIND improves 5.13% for

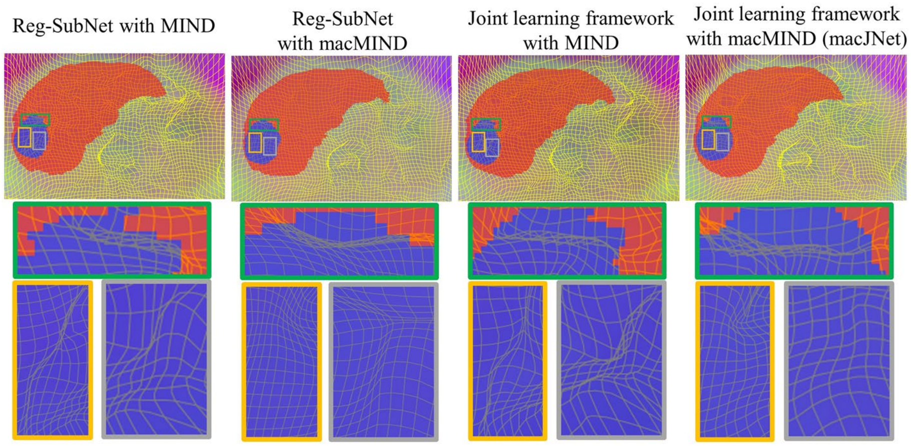

Reg-SubNet with MIND
Reg-SubNet with macMIND
Joint learning framework with MIND
Joint learning framework with macMIND (macJNet)

Fig. 4 Visualization of registration on a sample in the test dataset. The four columns show the deformation field results of Reg-SubNet-MIND, Reg-SubNet-macMIND, JNet-MIND and macJNet in sequence. The red region is the liver label and the blue region is the tumor label

Table 7 Registration results with different number of labels (mean ± std)

Amount	Liver			Image	
	TRE (mm)	DSC (%)	Hd_{95} (mm)	MI (%)	SSIM (%)
0%	4.90 ± 1.48	93.64 ± 1.07	5.39 ± 1.14	43.76 ± 4.83	53.00 ± 11.64
5%	4.93 ± 1.52	94.42 ± 0.85	4.94 ± 1.18	43.98 ± 4.70	53.72 ± 11.53
10%	5.18 ± 1.63	94.50 ± 0.92	4.80 ± 1.12	43.93 ± 4.60	54.41 ± 11.61
20%	5.24 ± 1.75	94.55 ± 0.91	4.63 ± 1.03	44.12 ± 4.58	54.85 ± 11.53
30%	**4.83 ± 1.49**	**94.75 ± 0.82**	**4.53 ± 1.11**	**44.12 ± 4.63**	**54.43 ± 11.62**
40%	4.79 ± 1.67	94.64 ± 0.82	4.52 ± 0.88	43.96 ± 4.55	54.22 ± 11.63
50%	5.05 ± 1.38	94.76 ± 0.85	4.46 ± 1.17	44.32 ± 4.64	54.39 ± 11.60
60%	5.10 ± 1.36	94.72 ± 0.89	4.59 ± 1.15	44.30 ± 4.71	54.05 ± 11.71
70%	5.27 ± 1.18	94.69 ± 0.86	4.53 ± 1.12	44.45 ± 4.68	54.24 ± 11.59
80%	5.31 ± 1.36	94.78 ± 0.79	4.40 ± 0.87	44.18 ± 4.72	54.17 ± 11.63
90%	4.97 ± 1.21	94.86 ± 0.89	4.20 ± 1.09	44.20 ± 4.52	53.94 ± 11.62
100%	5.10 ± 1.48	94.91 ± 0.81	4.21 ± 0.91	44.38 ± 4.65	54.09 ± 11.59

Bold values indicate better results than other methods

TRE, 3.58% for DSC, and 7.07% for Hd_{95} in tumor registration; improve 10.97% for TRE and 2.36% for Hd_{95} in liver registration.

To further explore the influence of weight of self-similarity context in dual-scales, the optimal ratios of α_1 and α_2 is verified in the macJNet. Figure 3 gives an overview of different ratios of α_1 and α_2, which also demonstrates that $\alpha_1/\alpha_2 = 7/3$ is an optimal ratio value for CT-MR liver registration. In addition, the change of TRE values also illustrates the effectiveness of multi-scale sampling in macMIND.

Evaluation of joint learning framework Organ labels of multimodal image pairs provide anatomical consistency constraint, which is considered as prior knowledge to guide alignment and deformation. However, manual labeling on multimodal images is a time-consuming task. Semi-supervised learning-based segmentation incorporated in a joint learning framework is a feasible way to provide segmentation labels for weakly-super-

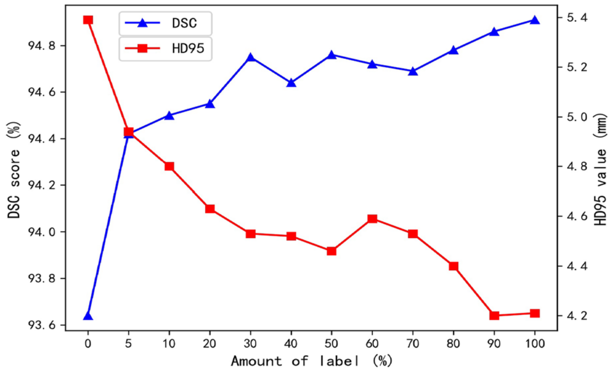


Fig. 5 The influence of different number of labels on macJNet

Table 8 Segmentation results on CT and MR with 30% labels (mean/std)

Modality	Methods	DSC (%)	Hd_{95} (mm)	Recall (%)	Precision (%)	RVD_{abs} (%)	VOE (%)
CT	Sub-SegNet	93.39 ± 3.31	19.15 ± 16.66	95.63 ± 6.24	91.50 ± 2.07	7.80 ± 4.25	12.24 ± 5.44
	macJNet	**95.38 ± 0.82**	**5.97 ± 3.21**	**96.05 ± 1.93**	**94.71 ± 1.24**	**2.53 ± 2.07**	**8.81 ± 2.19**
MR	Sub-SegNet	93.10 ± 2.39	20.95 ± 18.59	**95.23 ± 1.93**	91.20 ± 4.25	4.75 ± 6.22	12.81 ± 4.05
	macJNet	**94.12 ± 1.06**	**6.59 ± 2.14**	94.22 ± 2.13	**94.36 ± 2.72**	**3.92 ± 2.15**	**12.17 ± 1.86**

Bold values indicate better results than other methods

vised registration. In this experiment, macJNet and Reg-SubNet is performed to access the effectiveness of anatomical consistency constraint. Reg-SubNet is considered as unsupervised registration network here since there are no inputting segmentation labels into it. macJNet also used 30% labels to training the Seg-SubNet, and macMIND is used as metric in macJNet and Reg-SubNet. The performance of macJNet and Reg-SubNet is listed in Table 5. The statistical result shows that the label-based anatomical consistency plays an important role in organ boundary alignment. It significantly improves the liver registration performance in this experiment: improving 1.43% for TRE, 1.19% for DSC, 15.96% for Hd_{95}, 1.43% for SSIM. However, the influence of label-based anatomical consistency is diminished on the registration of internal lesion regions.

Some studies pointed out that the label-guided registration may receive diminishing or perturbing gradients [36, 37]. In the above experiment, DSC and Hd_{95} of the tumor are decreased due to the fact that the liver labels emphasize the alignment of the liver boundaries and ignores the physical properties of the deformation field, which yields some implausible deformation [31].

To investigate the effect of liver labels and modality-independent descriptors on the physical properties of the deformation field, the proportion of folding occurs (Jacobi determinant < 0) is calculated in different methods, as shown in Table 6. In the first set of experiments, the MIND descriptor and macMIND descriptor are separately applied to the Reg-SubNet (unsupervised registration). It is observed that macMIND performs significantly better than MIND with lower average proportion of folding points (0.63‱). In the second set of experiments, the two descriptors are applied separately to the joint learning framework (weakly-supervised registration), the average proportion of folding points in macMIND is also lower than that in MIND. It means that macMIND can effectively alleviates the negative impact of liver label and improve the physical properties of the deformation field. The visualization of deformation fields is shown in Fig. 4, which illustrates that macMIND effectively improves the physical properties of the deformation field.

Seg-SubNet is a semi-supervised segmentation network, which is influenced by the total amount of manual labeled images. To explore the influence of various amount of anatomy labels on registration, 0–100% different proportions of liver labels are input into Seg-SubNet by evaluating the registration metrics of liver registration. The results of liver registration are listed in Table 7. It obviously shows that the liver registration accuracy (DSC and H_{d95}) gradually increases with the increase of label amounts, which demonstrates that the anatomy consistency of multimodal images provides prior knowledge to guide registration. The anatomy labels play an important role in alignment of organ boundaries: liver registration would be significantly improved if very few labels (such as 5% labels) are input into the joint learning registration framework. 30% of total amount of label would be considered as a trade-off between the time-consuming manual label task and registration accuracy, which can be seen clearly in Fig. 5.

Multimodal image segmentation results

Although our study aims to improve the performance of multimodal deformable registration, macJNet also have an ability to improve the performance of multimodal image segmentation due to its multi-modality consistency constraint for segmentation labels. macJNet provides consistency between labels by mapping the moving label to the fixed label via a deformation field.

To quantitatively verify the improvement of segmentation of macJNet, DSC, Hd_{95}, recall, precision, absolute value of relative volume difference (RVD_{abs}) and volumetric overlap error (VOE) are used to evaluate the segmentation accuracy. RVD_{abs} and VOE are defined as:

$$\mathrm{RVD_{abs}} = \left| V_{\mathrm{seg}} / V_{\mathrm{gt}} - 1 \right| \times 100\%, \tag{3}$$

$$\mathrm{VOE} = \left(1 - \left(V_{\mathrm{seg}} \cap V_{\mathrm{gt}}\right) / \left(V_{\mathrm{seg}} \cup V_{\mathrm{gt}}\right)\right) \times 100\%, \tag{4}$$

where V_{seg} and V_{gt} symbol the segmentation volume and ground-truth volume, respectively.

The liver segmentation results on CT and MR images are listed in Table 8. Both macJNet and Sub-SegNet are trained with 30% labels. The statistical results of macJNet outperform the Sub-SegNet. Moreover, macJNet trained with 30% labels even surpasses the Sub-SegNet with 100% labels (DSC = 95.26, Hd_{95} = 8.71, Precision = 93.82, RVD_{abs} = 4.40, VOE = 9.04) for fixed (CT) image segmentation.

Conclusion

This article has proposed macJNet for multimodal image deformable registration. macJNet is a weakly-supervised multimodal image deformable registration network using joint learning framework and macMIND. The main advantage of macJNet is that it provides global sparse correspondences by semantic labels and local dense correspondences by macMIND, where macMIND provides the local modality-independent contextual information. macJNet consists of a registration network and two segmentation networks. Each segmentation network generates semantic anatomical labels as weakly-supervised information for registration; macMIND incorporates multi-orientation and multi-scale sampling patterns to build self-similarity context, which is modality-independent image structure features and used as dense local contextual information to guide the registration. The registration network also provides the consistency of anatomical labels by spatial mapping for segmentation networks to improve the performance of multimodal image segmentation. Experiments on 3D CT-MR liver images have been carried out to evaluate performance of macJNet. Experimental results indicate that our method achieves significant improvements in multimodal registration task.

In future studies, label-efficient deep learning methods will be incorporated into our method to further reduce the reliance on manually labeled images. In addition, the impact of sampling scale number and multi-scale information fusion ways on registration results will be investigated.

Methodology

Overview

In this work, macJNet is proposed to improve the accuracy of multi-modality image registration. macJNet is a weakly-supervised multimodal image deformable registration method, which incorporates two components: joint learning framework and macMIND. The joint learning framework is a single end-to-end architecture, which includes two segmentation networks and a registration network. Segmentation networks provide semantic anatomical labels for weakly-supervised registration by few-label learning; registration network improves the performance of segmentation by enforcing cross-modality consistency based on deformable spatial mapping. macMIND builds the local self-similarity context by multi-orientation and multi-scale sampling in a supporting

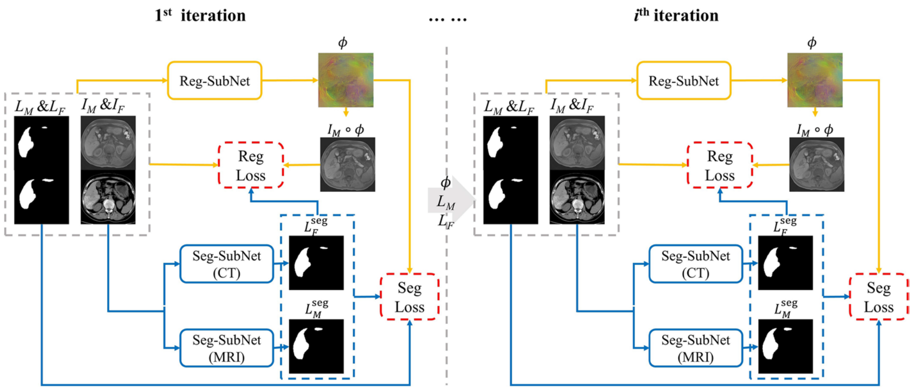


Fig. 6 Illustration of macJNet for CT-MRI registration. Image labels L comprise two subsets: manual annotation label subset L^{gt} as ground-truth in segmentation network, prediction label subset L^{seg} is generated by Seg-SubNets. $L_M = \{Lgt\ M, Lseg\ M\}$, $L_F = \{Lseg\ F, Lseg\ F\}$. For each iteration, Reg-SubNet takes I_M, I_F and their labels as input, outputs the deformation field φ, which provides the cross-modality consistency constrain for Seg-SubNets by mapping L_M to L_F. Seg-SubNets take I_M and I_F as input, and output $Lseg\ M$ and $Lseg\ F$ to provide semantic labels as anatomical prior knowledge for registration

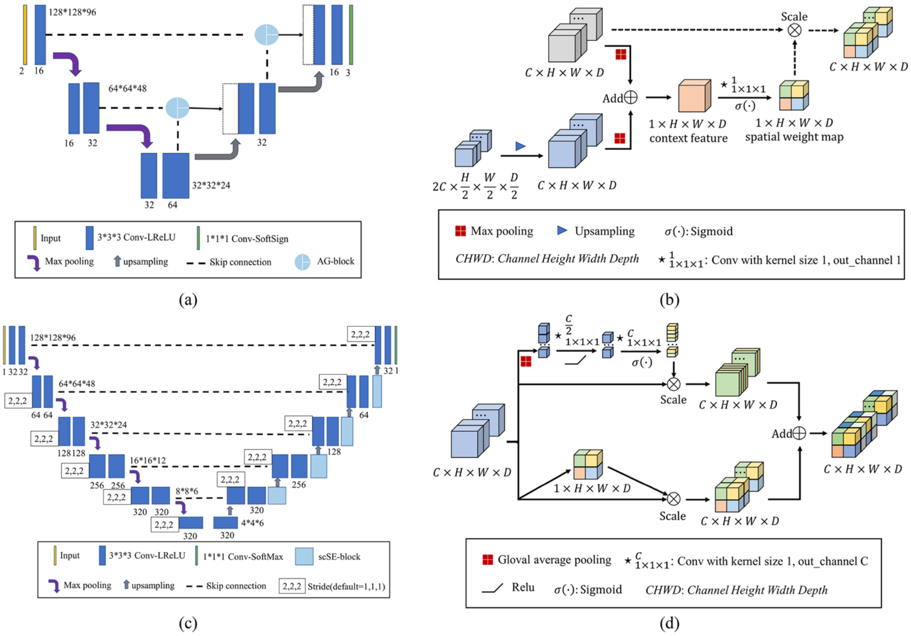


Fig. 7 Reg-SubNet and Seg-SubNet in the joint learning framework. **a** The architecture of Reg-SubNet; **b** AG-block in Reg-SubNet; **c** architecture of Seg-SubNet. **d** scSE-block in Seg-SubNet

window, which enriches the modality-independence contextual information to characterize cross-modality anatomical structures. Detail of the proposed method is described in "Joint learning framework" and "Reg-SubNet and Seg-SubNet".

Joint learning framework

macJNet comprises three sub-networks, a weakly-supervised registration sub-network (Reg-SubNet) and two semi-supervised segmentation sub-network (Seg-SubNet) for dual-modality image segmentation. K_{u} unlabeled multimodal image pairs and K_{l} labeled image pairs ($K_{\mathrm{u}} > K_{\mathrm{l}}$) are input into the network to optimize macJNet. Specifically, an alternately update strategy is used to optimize Reg-SubNet and Seg-SubNets in the joint learning framework. In the registration update stage, I_F, I_M and their anatomy labels (including K_{l} pairs with manual labels L^{gt} and K_{u} pairs with segmentation labels L^{seg} created by Seg-SubNets) are input into Reg-SubNet to optimize the dense deformation fields ϕ. In the segmentation update stage, K_{u} unlabeled image pairs and K_{l} labeled image pairs are input into the Seg-SubNets to generate the segmentation labels ($Lseg\ M$ and $Lseg\ F$), where the dense deformation fields created by Reg-SubNet maps $Lseg\ M$ to $Lseg\ F$ for cross-modality consistency constraint. The overview of the joint learning framework is illustrated in Fig. 6.

The main advantages of joint learning in macJNet are as follows: (1) incorporating two correlated tasks in a single framework to improve the performance of registration; (2) allowing to use existing task-specific networks for registration and segmentation. It is noteworthy that our work does not focus on the design of an elaborate registration network. The main aim of this work is to propose a general framework for weakly-supervised registration, any task-specific registration or segmentation networks could be used

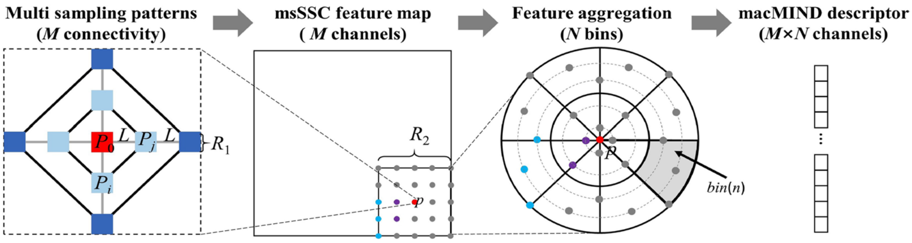


Fig. 8. 2D illustration of cascaded feature calculations of macMIND. The "Multi sampling partners" sketch illustrates the msSSC with multi-sampling patterns (multi-scale sampling and multi-orientation sampling). msSSC includes some different scale self-similarity contexts (SSC). The left sketch illustrates a dual-scale SSC: the small-scale SSC includes the central patch P_0 (red box) and its closer 4-neighborhood (light blue boxes); the larger-scale SSC includes the central patch P_0 and its farther 4-neighborhood (dark blue boxes). Each SSC includes more connectivity (black lines and gray line) than MIND (gray lines), which leads macMIND to incorporate more orientation sampling. L and R_1 symbolize the patch distance and size, respectively. The msSSC feature map with M channels are created. The "feature aggregation" sketch shows the N bins (here $N = 16$) in log-polar space with 8-angle intervals and 2-radial intervals. One of the bins is colored with gray. macMIND translates each voxel in an image to a $M \times N$ matrix by macMIND. Finally, macMIND feature map is created as a $M \times N$ channel image for registration

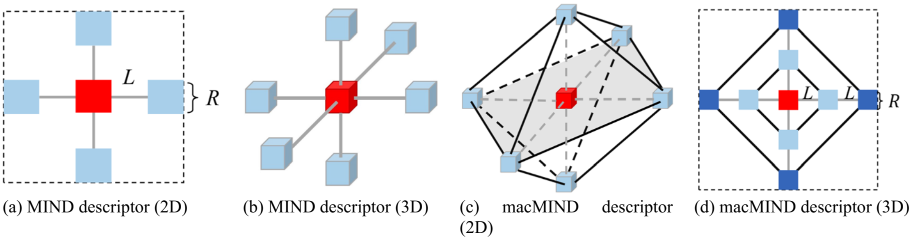


Fig. 9 Illustration of MIND and macMIND descriptor. **a** The dotted box illustrates the supporting window, the red box illustrates the central patch of the supporting window, its 6-neighborhood patches are colored in blue. L and R symbolize the patch distance and size, respectively. **b** The 3D structure of 6-connectivities (3-orientation) in MIND. **c** The single scale 3D structure layout of 18-connectivities (9-orientation) in macMIND. The black (shown as black solid lines and black dotted lines) 12-connectivities are introduced by macMIND, while the gray 6-connectivities (3-orientation) attached to MIND. **d** The supporting window of macMIND similar with MIND. Light blue and dark blue patches indicate the dual sampling scales, and black and gray connections indicate the multi sampling orientation

as sub-networks in this framework. Some other works [27, 38] joint the registration and segmentation through multi-task learning. Multi-task learning methods joint the two tasks using hard or soft parameter sharing, which needs to change the architecture of existing networks.

Reg-SubNet and Seg-SubNet

In this study, LapIRN is adopted to build Reg-SubNet (shown in Fig. 7a). LapIRN [35] is a deep Laplacian pyramid image registration network with a 3D UNet-like architecture [39] and mitigates the large-deformation problem via a coarse-to fine scheme [35]. AG-blocks [40] (shown in Fig. 7b) are added into the LapIRN to filter the features by propagating through the skip connections. AG-blocks employ multi-level spatial and contextual information to highlight the regions with large discrepancies. The nnUNet

[41] is applied to build Seg-SubNet (shown in Fig. 7c) due to its excellent performance in medical image segmentation. scSE-blocks [42] (shown in Fig. 7d) are added into the decoding layers to suppress insignificant information in both spatial and channel dimensions. In the training stage, Dice loss is used to measure the similarity in label pairs.

Multi-sampling cascaded modality independent neighborhood descriptor

Modality independent neighborhood descriptor

MIND is a well-known image representor [20] for multi-modality image registration, which represents local self-similarity structures by calculating the difference between patches within a local neighborhood. For any point x in image I, the MIND feature can be represented by Gaussian kernel distance between center point x and its 6-neighborhood patches, as shown in Fig. 9b. Assuming that the n-th patch in the 6-neighborhood centered at x_n, MIND can be expressed as:

$$\mathrm{MIND}(I, x, x_n) = \exp\left(-\frac{D_p(I, x, x_n)}{V(I, x)}\right) \tag{5}$$

where $D_p(I, x, x_n)$ donates the mean squared difference between two patches, which, respectively, locate at x and x_n. P is defined as the set of displacements from any voxel in a patch to the center of the patch.

$$D_p(I, x, x_n) = \frac{1}{|P|}\sum_{t \in P}(I(x+t) - I(x_n+t))^2, \tag{6}$$

$V(I, x)$ is an estimation of the local variance, defined as the expectation of D_p:

$$V(I, x) = \frac{1}{6}\sum_{n=1}^{6} D_p(I, x, x_n). \tag{7}$$

However, MIND computes self-similarity between the center patch and its 6-neighborhood patches with the simple sampling pattern (shown in Fig. 9), which cannot handle the large deformation and high complex dense correspondence.

macMIND

Inspired by MIND, a multi-sampling cascaded modality independent neighborhood descriptor (macMIND) is proposed to improve performance of multimodal image deformable registration. The motivation of macMIND is to incorporate more abundant sampling patterns for representing the complex cross-modality structure features, which contributes to find dense correspondence in multimodal images.

The macMIND descriptor incorporates cascaded feature calculations: (1) multi-scale self-similarity context (msSSC) feature map calculation with multi-sampling patterns; (2) feature aggregation in 3D log-polar bins. Figure 8 illustrates the cascaded feature calculations of macMIND. macMIND extracts the $M \times N$-channels feature map of every voxel in the image. The specific implementation process and its advantages will be detailed in the following sections.

Multi-sampling patterns of msSSC

Multi-sampling patterns (multi-orientation sampling and multi-scale sampling) are introduced to encode the self-similarity context, which is robust and accurate cross-modality feature representation. Specifically, given a certain patch layout P_Ω, the central patch P_0 of size $R_1 \times R_1 \times R_1$ centered at voxel p and the distance between P_0 and its 6-neighborhood patches is L (Fig. 8a). The self-similarity context $SSC(I, P_\Omega)$ is defined as:

$$\mathrm{SSC}(I, P_\Omega) = \sum_{P_i, P_j \in P_\Omega} \exp\left(- \frac{\|e_P\| SSD\left(I, P_i, P_j\right)}{\sum\limits_{P_i, P_j \in P_\Omega} SSD\left(I, P_i, P_j\right)} \right), \tag{8}$$

where I is an image, $P_\Omega = \{P_0, P_1, \ldots, P_6\}$, P_i and P_j are the symbols of arbitrary patches in P_Ω, $\|e_P\|$ denotes the total number of patch connections. $SSD(I, P_i, P_j)$ denotes the sum of squared difference between patch P_i and P_j, which is formulated as:

$$\mathrm{SSD}\left(I, P_i, P_j\right) = \frac{1}{\|P\|}\left(I(P_i) - I\left(P_j\right)\right)^2, \tag{9}$$

where $\|P\|$ represents the total number of voxels in patch P. In Eq. (8), SSC is computed as the sum of squared difference between two patches to represent the self-similarity. As shown in Fig. 9d, there are 18-connectivity in SSC within single scale, which can be divided into 9 orientations. Therefore, a multi-orientation sampling pattern is introduced into macMIND, which leads macMIND to represent the complex deformations.

The multi-scales self-similarity context (msSSC) is further computed to represent the large deformation (large geometrical variations) in the multimodal images. msSSC can be reformulated as:

$$\mathrm{msSSC}(p) = \sum_{k=1}^{K} \alpha_k \mathrm{SSC}\left(I, P_\Omega^k\right) = \sum_{k=1}^{K} \alpha_k \sum_{P_i, P_j \in P_\Omega^k} \exp\left(- \frac{\|e_P\| \mathrm{SSD}\left(I, P_i^k, P_j^k\right)}{\sum\limits_{P_i, P_j \in P_\Omega^k} \mathrm{SSD}\left(I, P_i^k, P_j^k\right)} \right), \tag{10}$$

where K is the total number of scales, and α_k denotes the weight of multi-self-similarity context in the k-th scale. The Eq. (10) formulates the multi-self-similarity context in a weighted sum way. It is a simple way to fuse the multi-scale information in consideration

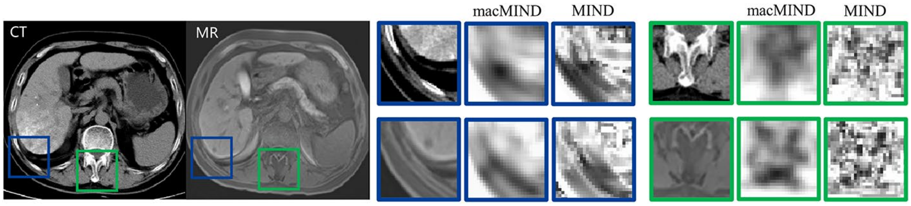
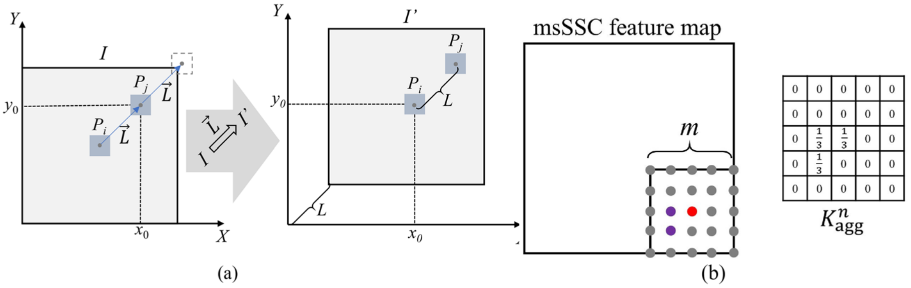


Fig. 10 Feature map visualization of macMIND and MIND. macMIND shows its advantage for representing complex anatomical structures. The multi-channel feature map is translated to a single-channel image by calculating the average value of channel-dimension for visualization. In this figure, macMIND is calculated with $K = 2$, $\alpha_1 = 0.7$, $\alpha_2 = 0.3$



0	0	0	0	0
0	0	0	0	0
0	$\frac{1}{3}$	$\frac{1}{3}$	0	0
0	$\frac{1}{3}$	0	0	0
0	0	0	0	0

Fig. 11 The convolution operations in SSD calculation and feature aggregation. **a** SSC feature map is calculated in the manner of convolution operations. P_i is the center patch of the image, P_j is a neighborhood patch of P_i. P_i would overlap P_j by shifting with $\overrightarrow{L}$. This operation translates the computation of SSD to a voxel-wise squared difference. **b** Feature aggregation with a convolution operation. The kernels are designed according to the spatial distribution of voxels in bins

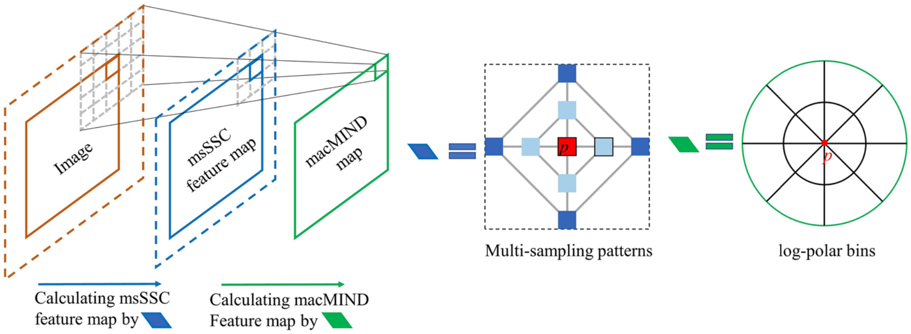


Fig. 12 Illustration of cascaded convolution operations in macMIND

of the limitations of computing resources. The other way is to concatenate $SSC(I, P_{\Omega})$ of each scale along the channel dimension.

In summary, self-similarity in MIND is calculated based on the central patch (shown in Fig. 9a, b), which has the disadvantage that the noise in central patch takes adverse effect on the self-similarity. Compared with MIND, msSSC has two advantages: (1) multi-orientation sampling: utilizing all pairwise connectivity (18- connectivity) within central patch and its 6-neighbourhood to build a 9-orientation sampling pattern (shown in Fig. 9c); (2) multi-scale sampling: incorporating multi-scale self-similarity context in a supporting window (shown in Fig. 9c). The multi-sampling pattern in msSSC leads macMIND to represent the complex and large deformation in multimodal images.

Feature aggregation in 3D log-polar bins

Each point in the msSSC feature map is aggregated into the log-polar bins [24] to robustly represent the cross-modality structural information in deformable registration [43, 44]. A patch with size $R_2 \times R_2 \times R_2$ and central at voxel p on msSSC is selected, and all voxels in the patch are transformed into local 3D log-polar space. The 3D log-polar space is divided into N bins based on N_a angle intervals, N_r radial intervals and N_h height intervals ($N = N_a \times N_r \times N_h$). The values in each bin are calculated and the average values are concatenated into a $M \times N$-dimension vector as a macMIND descriptor. The macMIND is defined as

$$\text{macMIND}(I,p) = \underset{n\in\{1,2\cdots N\}}{\text{cat}} \left(\underset{p\in bin(n)}{\text{average}} \left(\text{msSSC}(p)\right) \right), \tag{11}$$

where 'cat' symbolizes the vector concatenation. Finally, macMIND translates each voxel p in image to a $M \times N$ vector. Here, the msSSC feature image is aggregated by utilizing average pooling in bins instead of max-pooling to maintain the fine-scale matching details [45].

Figure 10 shows the comparisons of feature map between macMIND and MIND on CT-MRI images. Two typical locations with different image structures are selected: (1) the boundary between liver and abdomen (blue window), it is a latent region with large deformation; (2) spine (green window), it is a region with complex structures. It is obviously observed that the macMIND feature map accurately represents the modality-independent features (e.g., tissue boundary and shape), and is more continuous and smoother than MIND.

Cascaded feature extraction

Actually, SSD is computed and feature map aggregation is performed in a convolutional manner due to its computational efficiency. Specifically, for computing $SSD(I, P_i, P_j)$, image I' is obtained by shifting the image I by a vector $\overrightarrow{L}$, as shown in Fig. 11a. $I(P_j)$ is equal to $I'(P_i)$ since the distance between patch P_i and patch P_j is $\left\|\overrightarrow{L}\right\|$. Then, the voxel-wise squared difference is calculated in the minus manner of I and I': $I(P_i)$-$I(P_j) = I(P_i)$-$I'(P_i)$. Finally, the patch-wise squared difference can be obtained from voxel-wise squared difference by convolution with a $R_1 \times R_1 \times R_1$ sized kernel K_{SSD}. K_{SSD} is designed as an average pooling kernel. The $SSD(I, P_i, P_j)$ computation in Eq. (9) can be effectively substitute, which is reformulated as:

$$\text{SSD}\left(I, P_i, P_j\right) = K_{\text{SSD}} \otimes (I - I')^2, \tag{12}$$

where $\otimes$ is the convolution operator. For aggregating the point of msSSC feature map in 3D log-polar bins, a specific convolution kernel Kn agg (n is the scale parameter in msSSC, $n = 1,2,\ldots,N$) with size of $R_2 \times R_2 \times R_2$ is designed on the msSSC feature map (shown in Fig. 11b) for each bin. Kn agg transforms the mean value calculation to a convolution operation. The Eq. (11) could be reformulated as

$$\text{macMIND}(I,p) = \underset{n\in\{1,2\cdots N\}}{\text{cat}} \left(K^n_{\text{agg}} \otimes \text{SSC}(p) \right), \tag{13}$$

The cascade convolution operations of macMIND are similar to the feature learning in two consecutive encoder layers of CNN (shown in Fig. 12), which has two advantages for registration: representing deeper and more complex features, enlarging the receptive field of feature representation with low computational cost [24].

The sampling density is a main difference between MIND [21], macMIND, and DSC [24]. On one hand, compared with MIND, macMIND increases the sampling density by utilizing all connectivity of patch layout to encode comprehensive information in SSC feature map. All connectivity of patch layout also introduces multi-scale and multi-orientation sampling patterns. The increase of sampling density enriches the modality-independence contextual information for dense correspondence cross multimodal images. On the other hand, in comparison to the deep learning-based sampling on the self-similarity surface [24], macMIND supplies a sparse sampling with the fixed patterns. Although some dense sampling patterns have been proposed to build more elaborate cross-modality descriptors (such as DSC [24] and DASC [45]), they would be computationally intractable for 3D medical images. The sparse sampling patterns are necessary for efficient computation in 3D medical image registration. The patch layout in the supporting window of macMIND is much sparser than the dense self-correlation surfaces in DSC and DASC.

Loss function in joint learning framework

Dual-similarity-based loss is proposed for registration: a macMIND-based similarity metric to capture the dense correspondence of modality-independent texture characteristics, DSC-based similarity to capture the semantic consistency of anatomical characteristics in multimodal images. DSC value of label images is used as loss to guide the Seg-SubNet training.

Dual similarity-based loss for multimodal image registration

The loss function for Reg-SubNet is defined as: $E_{\text{reg}}=E_{sim}(I_F,\ I_M\circ\phi)+\lambda_{\text{label}}E_{\text{label}}(L_F,\ L_M\circ\phi)+\lambda_{\text{smo}}E_{\text{smo}}(\phi)$. $E_{\text{sim}}(I_F,\ I_M\circ\phi)$ takes the form as

$$E_{\text{sim}}(I_F, I_M \circ \phi) = \frac{1}{|\Omega^3|}\sum_{p\in\Omega^3}(\text{macMIND}(I_F,p) - \text{macMIND}(I_M \circ \phi, p))^2, \tag{14}$$

E_{sim} measures the local difference between the pair of macMIND maps, $|\Omega^3|$ is the total voxel number of the image. E_{label} measures the DSC value of fixed label image and warped label image. In addition, a diffusion regularization on the spatial gradients of ϕ is added to encourage a smooth deformation field, which is defined as

$$E_{\text{smo}}(\phi)=\sum_{p\in\Omega^3}\|\nabla\phi(p)\|^2. \tag{15}$$

Loss function for semi-supervised segmentation

The loss function for Seg-SubNet is defined as:

$$\begin{cases} E_{\text{seg}} = E_{\text{DSC}}\left(L_F^{\text{seg}}, L_M^{\text{seg}} \circ \phi\right) & L^{\text{gt}} \text{ not existed} \\ E_{\text{seg}} = E_{\text{DSC}}\left(L_F^{\text{seg}}, L_F^{\text{gt}}\right) + E_{\text{DSC}}\left(L_M^{\text{seg}}, L_M^{\text{gt}}\right) & L^{\text{gt}} \text{ existed} \end{cases} \tag{16}$$

L^{seg} represents the segmentation label output by Seg-SubNet, L^{gt} represents the ground-truth. $E_{\text{DSC}}(L^{\text{seg}}, L^{\text{gt}})$ guides the segmentation results to the ground-truth, and E_{DSC}(Lseg F, Lseg $M\circ\phi$) guides different Seg-SubNets to generate consistent segmentations labels.

Author contributions

ZZ and BH wrote the main manuscript; XQ revised the manuscript; JH and MS prepared figures; JJ provided and labeled the medical images; YD and ZZ acquired the financial support for the project leading to this publication.

Funding

This work was supported partly by National Key Technology Research Development Program (2018YFA0703101), National Natural Science Foundation of China (61971413, 62271480), Youth Innovation Promotion Association CAS (2021324), Jiangsu Key Technology Research Development Program (BE2022049-2, BE2021612, 2021012-5), Science and Technology Development Project of Suzhou (SKY2022052), Jiangsu International Cooperation Project (BZ2022028), Key Research and Development Program of Shandong (2019GHZ002).

Availability of data and materials

Not applicable.

Declarations

Ethics approval and consent to participate

All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Received: 10 January 2023 Accepted: 27 July 2023
Published online: 19 September 2023

References

1. Dosovitskiy A, Fischer P, Ilg E, Hausser P, Hazirbas C, Golkov V, van der Smagt P,. Cremers D, Brox T. FlowNet: Learning Optical Flow with Convolutional Networks. In: 2015 IEEE International Conference on Computer Vision (Iccv), pp. 2758–2766; 2015.
2. Sokooti H, de Vos B, Berendsen F, Lelieveldt BPF, Išgum I, Staring M. Nonrigid Image Registration Using Multi-scale 3D Convolutional Neural Networks. Medical Image Computing and Computer Assisted Intervention—MICCAI 2017. pp. 232–239.
3. Yang X, Kwitt R, Styner M, Niethammer M. Quicksilver: fast predictive image registration—a deep learning approach. Neuroimage. 2017;158:378–96.
4. Rohé M-M, Datar M, Heimann T, Sermesant M, Pennec X. SVF-Net: Learning Deformable Image Registration Using Shape Matching. In: Medical Image Computing and Computer Assisted Intervention—MICCAI 2017. pp. 266–274.
5. Fan JF, Cao XH, Yap EA, Shen DG. BIRNet: brain image registration using dual-supervised fully convolutional networks. Med Image Anal. 2019;54:193–206.
6. Cao X, Yang J, Zhang J, Nie D, Kim M, Wang Q, Shen D. Deformable Image Registration Based on Similarity-Steered CNN Regression. Medical Image Computing and Computer Assisted Intervention—MICCAI 2017. pp. 300–308.
7. Avants BB, Epstein CL, Grossman M, Gee JC. Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain. Med Image Anal. 2008;12(1):26–41.
8. Vercauteren T, Pennec X, Perchant A, Ayache N. Diffeomorphic demons: efficient non-parametric image registration. Neuroimage. 2009;45(1):S61–72.
9. Haskins G, Kruger U, Yan PK. Deep learning in medical image registration: a survey. Mach Vis Appl 2020;31(1).
10. de Vos BD, Berendsen FF, Viergever MA, Sokooti H, Staring M, Isgum I. A deep learning framework for unsupervised affine and deformable image registration. Med Image Anal. 2019;52:128–43.
11. Balakrishnan G, Zhao A, Sabuncu MR, Guttag J, Dalca AV. VoxelMorph: a learning framework for deformable medical image registration. IEEE Trans Med Imaging. 2019;38(8):1788–800.
12. Mok TCW, Chung ACS. Fast Symmetric Diffeomorphic Image Registration with Convolutional Neural Networks. In: 2020 IEEE/Cvf Conference on Computer Vision and Pattern Recognition (Cvpr), pp. 4643–4652, 2020.
13. Wang J, Zhang MM. DeepFLASH: an efficient network for learning-based medical image registration. In: 2020 IEEE/Cvf Conference on Computer Vision and Pattern Recognition (Cvpr), pp. 4443–4451, 2020.
14. Yan PK, Xu S, Rastinehad AR, Wood BJ. Adversarial image registration with application for MR and TRUS image fusion. In: Machine Learning in Medical Imaging: 9th International Workshop, Mlmi 2018; 11046:197–204
15. Kim S, Min D, Ham B, Lin S, Sohn K. FCSS: fully convolutional self-similarity for dense semantic correspondence. IEEE Trans Pattern Anal Mach Intell. 2019;41(3):581–95.
16. Mahapatra D, Antony B, Sedai S, Garnavi R. Deformable medical image registration using generative adversarial networks. In: 2018 IEEE 15th International Symposium on Biomedical Imaging (Isbi 2018), pp. 1449–1453, 2018.
17. Fan JF, Cao XH, Wang Q, Yap PT, Shen DG. Adversarial learning for mono- or multi-modal registration. Med Image Anal. 2019;58:101545.
18. Xu Z, Luo J, Yan J, Pulya R, Li X, Wells W 3rd, Jagadeesan J. Adversarial uni- and multi-modal stream networks for multimodal image registration. Med Image Comput Comput Assist Interv. 2020;12263:222–32.

19. Farnia F, Ozdaglar A. Do GANs always have Nash equilibria?. In: International Conference on Machine Learning. 2020; 119.
20. Shechtman E, Irani M. Matching local self-similarities across images and videos. In: 2007 IEEE Conference on Computer Vision and Pattern Recognition, vol. 1–8, pp. 1744; 2007.
21. Heinrich MP, Jenkinson M, Bhushan M, Matin T, Gleeson FV, Brady SM, Schnabel JA. MIND: modality independent neighbourhood descriptor for multi-modal deformable registration. Med Image Anal. 2012;16(7):1423–35.
22. Torabi A, Bilodeau GA. Local self-similarity-based registration of human ROIs in pairs of stereo thermal-visible videos. Pattern Recogn. 2013;46(2):578–89.
23. Ye YX, Shan J. A local descriptor based registration method for multispectral remote sensing images with non-linear intensity differences. ISPRS J Photogramm Remote Sens. 2014;90:83–95.
24. Kim S, Min D, Lin S, Sohn K. Dense cross-modal correspondence estimation with the deep self-correlation descriptor. In: IEEE Transactions on Pattern Analysis and Machine Intelligence. 2021; 43(7): 2345–2359.
25. Xu ZL, Niethammer M. DeepAtlas: Joint Semi-supervised Learning of Image Registration and Segmentation. In: Medical Image Computing and Computer Assisted Intervention—Miccai 2019, Pt Ii, vol. 11765, pp. 420-429, 2019.
26. Mahapatra D, Ge ZY, Sedai S, Chakravorty R. Joint Registration And Segmentation Of Xray Images Using Generative Adversarial Networks. In: Machine Learning in Medical Imaging: 9th International Workshop, Mlmi 2018, vol. 11046, pp. 73–80, 2018.
27. Estienne T, Vakalopoulou M, Christodoulidis S, Battistela E, Lerousseau M, Carre A, Klausner G, Sun R, Robert C, Mougiakakou S, Paragios N, Deutsch E. U-ReSNet: ultimate coupling of registration and segmentation with deep nets. In: Medical image computing and computer assisted intervention—Miccai 2019, Pt Iii, vol. 11766, pp. 310–319; 2019.
28. Shao W, Bhattacharya I, Soerensen SJC, Kunder CA, Wang JB, Fan RE, Ghanouni P, Brooks JD, Sonn GA, Rusu M. Weakly Supervised Registration of Prostate MRI and Histopathology Images. In: Medical Image Computing and Computer Assisted Intervention - Miccai 2021, Pt Iv, vol. 12904, pp. 98–107, 2021.
29. Blendowski M, Hansen L, Heinrich MP. Weakly-supervised learning of multi-modal features for regularised iterative descent in 3D image registration. Med Image Anal. 2021;67:101822.
30. Elmahdy MS, Wolterink JM, Sokooti H, Isgum I, Staring M. Adversarial optimization for joint registration and segmentation in prostate CT radiotherapy. In: Medical Image Computing and Computer Assisted Intervention—Miccai 2019, Pt Vi, vol. 11769, pp. 366–374, 2019.
31. Hu YP, Modat M, Gibson E, Ghavami N, Bonmati E, Moore CM, Emberton M, Noble JA, Barratt DC, Vercauteren T, Label-Driven weakly-supervised learning for multimodal deformable image registration. In: 2018 IEEE 15th International Symposium on Biomedical Imaging (Isbi 2018), pp. 1070–1074, 2018.
32. Wang Z, Bovik AC, Sheikh HR, Simoncelli EP. Image quality assessment: from error visibility to structural similarity. IEEE Trans Image Process. 2004;13(4):600–12.
33. Lian CY, Li XM, Kong LK, Wang JC, Zhang W, Huang XY, Wang LS. CoCycleReg: collaborative cycle-consistency method for multi-modal medical image registration. Neurocomputing. 2022;500:799–808.
34. Klein S, Staring M, Murphy K, Viergever MA, Pluim JPW. elastix: a toolbox for intensity-based medical image registration. IEEE Trans Med Imaging. 2010;29(1):196–205.
35. Mok TCW, Chung ACS. Large deformation diffeomorphic image registration with laplacian pyramid networks. In: Medical Image Computing and Computer Assisted Intervention—MICCAI 2020, pp. 211–221, 2020.
36. Qiu L, Ren HL. RSegNet: a joint learning framework for deformable registration and segmentation. IEEE Trans Autom Sci Eng. 2022;19(3):2499–513.
37. Qiu L, Ren HL. U-RSNet: an unsupervised probabilistic model for joint registration and segmentation. Neurocomputing. 2021;450:264–74.
38. Elmahdy MS, Beljaards L, Yousefi S, Sokooti H, Verbeek F, Van der Heide UA, Staring M. Joint registration and segmentation via multi-task learning for adaptive radiotherapy of prostate cancer. IEEE Access. 2021;9:95551–68.
39. Çiçek Ö, Abdulkadir A, Lienkamp SS, Brox T, Ronneberger O. 3D U-Net: learning dense volumetric segmentation from sparse annotation. In: Medical Image Computing and Computer-Assisted Intervention—MICCAI 2016. pp. 424–432.
40. Oktay O, Schlemper J, Folgoc LL, Lee MJ, Heinrich MP, Misawa K, Mori K, McDonagh SG, Hammerla NY, Kainz B, Glocker B, Rueckert DJA, Attention U-Net: learning where to look for the pancreas. arXiv preprint, vol. arXiv:1804.03999, 2018.
41. Isensee F, Jaeger PF, Kohl SAA, Petersen J, Maier-Hein KH. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nat Methods. 2021;18(2):203–11.
42. Roy AG, Navab N, Wachinger C. Concurrent Spatial and Channel 'Squeeze & Excitation' in Fully Convolutional Networks. In: Medical Image Computing and Computer Assisted Intervention—Miccai 2018, Pt I, vol. 11070, pp. 421–429, 2018.
43. Calonder M, Lepetit V, Ozuysal M, Trzcinski T, Strecha C, Fua P. BRIEF: computing a local binary descriptor very fast. IEEE Trans Pattern Anal Mach Intell. 2012;34(7):1281–98.
44. Chatfield K, Philbin J, Zisserman A. Efficient retrieval of deformable shape classes using local self-similarities. In: 2009 IEEE 12th International Conference on Computer Vision Workshops (ICCV Workshops). [v.1], Kyoto, Japan, 2009, pp. 264–271.
45. Kim S, Min D, Ham B, Do MN, Sohn K. DASC: robust dense descriptor for multi-modal and multi-spectral correspondence estimation. IEEE Trans Pattern Anal Mach Intell. 2017;39(9):1712–29.

Publisher's Note