



Residual Aligner-based Network (RAN): Motion-separable structure for coarse-to-fine discontinuous deformable registration

Jian-Qing Zheng^{a,*}, Ziyang Wang^c, Baoru Huang^d, Ngee Han Lim^a, Bartłomiej W. Papież^b

^a The Kennedy Institute of Rheumatology, University of Oxford, UK

^b Big Data Institute, University of Oxford, UK

^c Department of Computer Science, University of Oxford, Oxford, UK

^d The Hamlyn Centre for Robotic Surgery, Imperial College, London, UK

ARTICLE INFO

Dataset link: <https://zenodo.org/record/3835682>, https://github.com/ucl-candi/datasets_deep_demo/archive/abdct.zip

Keywords:

Discontinuous deformable registration
Motion-separable structure
Motion disentanglement
Coarse-to-fine registration

ABSTRACT

Deformable image registration, the estimation of the spatial transformation between different images, is an important task in medical imaging. Deep learning techniques have been shown to perform 3D image registration efficiently. However, current registration strategies often only focus on the deformation smoothness, which leads to the ignorance of complicated motion patterns (e.g., separate or sliding motions), especially for the intersection of organs. Thus, the performance when dealing with the discontinuous motions of multiple nearby objects is limited, causing undesired predictive outcomes in clinical usage, such as misidentification and mislocalization of lesions or other abnormalities. Consequently, we proposed a novel registration method to address this issue: a new Motion Separable backbone is exploited to capture the separate motion, with a theoretical analysis of the upper bound of the motions' discontinuity provided. In addition, a novel Residual Aligner module was used to disentangle and refine the predicted motions across the multiple neighboring objects/organs. We evaluate our method, Residual Aligner-based Network (RAN), on abdominal Computed Tomography (CT) scans and it has shown to achieve one of the most accurate unsupervised inter-subject registration for the 9 organs, with the highest-ranked registration of the veins (Dice Similarity Coefficient (%)/Average surface distance (mm): 62%/4.9mm for the vena cava and 34%/7.9mm for the portal and splenic vein), with a smaller model structure and less computation compared to state-of-the-art methods. Furthermore, when applied to lung CT, the RAN achieves comparable results to the best-ranked networks (94%/3.0mm), also with fewer parameters and less computation.

1. Introduction

Alignment of multiple images, also known as image registration (Sotiras et al., 2013), is a crucial task in medical image analysis applications. In medical imaging, it allows for comparison across multiple acquisitions over time for longitudinal analysis (intra-subject registration), between different types of scanners for integration and correlation of information from various modalities (multi-modal registration), and between different individuals for population-specific studies and group-level statistics (inter-subject registration).

Image registration can be defined as the estimation of the spatial transformation $\phi: \mathbb{R}^n \rightarrow \mathbb{R}^n$, represented by corresponding parameters for linear spatial transform (such as rigid/affine transform) or a series of motions (or displacements) for deformation denoted by $\phi[x] \in \mathbb{R}^d$ at the coordinate $x \in \mathbb{Z}^d$ of a target image $I^t \in \mathbb{R}^n$ from a source image $I^s \in \mathbb{R}^n$, where n is the size of a 3D image defined as $n = H \times W \times T$, and d, T, H, W denoting the image dimension, thickness, height, and

width, respectively. Originally, image registration was solved as an optimization problem by minimization of a dissimilarity metric D and a regularization term S :

$$\tilde{\phi} = \underset{\phi}{\operatorname{argmin}} (D(\phi(I^s), I^t) + \lambda S(\phi, I^t)) \quad (1)$$

where $\tilde{\phi}$ denotes the estimated spatial transform, λ denotes the weight of the regularization. Several methods including Demons (Thirion, 1998) or Free Form Deformations (Rueckert et al., 1999) have been proposed to solve Eq. (1), however, they are likely to get trapped in the local optimum and their inference performance and efficiency are limited due to iterative optimization of highly dimensional, non-convex problem (Fischer and Modersitzki, 2008).

More recently, the registration is performed via (convolutional) neural networks \mathcal{R} using the feature maps $F^s, F^t \in \mathbb{R}^{c \times n}$ extracted from I^s and I^t respectively, (and c denotes the number of feature channels)

* Corresponding author.

E-mail addresses: jianqing.zheng@kennedy.ox.ac.uk (J.-Q. Zheng), bartlomiej.papiez@bdi.ox.ac.uk (B.W. Papież).

by directly regressing the spatial transformation (Balakrishnan et al., 2018; Mok and Chung, 2020) with one-attempt registration (contrast to iterative or progressive registration):

$$\phi = \mathcal{R}(F^s, F^t; \mathbf{w}) \quad (2)$$

with the training process based on minimizing the loss function (e.g. given in Eq. (1)) with the trainable weights \mathbf{w} (\mathbf{w} are omitted in the remaining part of the paper to simplify the notation), which is called Direct Regression (DR) Registration (see Fig. 1). However, the DR methods fail in dealing with large or complex motions such as sliding motion due to the limited capture range of the receptive field of convolution layers.

To address the limited capture range problem, Attention-based (Attn) mechanism (Vaswani et al., 2017) can be used for feature correspondence or alignment in Li et al. (2021), Sun et al. (2021), Heinrich (2019), Zheng et al. (2022, 2023) and Chen et al. (2022, 2021a) to obtain the global receptive field, with so-called attention score matrices to quantify the correspondence between each pair of pixels from two images. In this approach, each element of the feature map F^t , represented as a key vector, is compared with the query vectors from F^s in an attention score matrix $\Phi \in \mathbb{R}^{n \times n}$. The alignment by single-head cross attention can be formulated as:

$$\begin{cases} \Phi = \text{softmax}(C^q(F^s)^\top C^k(F^t)) \\ \phi(F^s) := F^s \Phi \end{cases} \quad (3)$$

where C^q and C^k denote the linear transformation for query and key feature vectors. The usage of multiple attention score matrices $\Phi = (\Phi_1, \Phi_2, \dots, \Phi_m)$ is called M-H attention. However, the calculation of the attention score matrix with $\mathcal{O}(n^2)$ complexity leads to large memory consumption and computational inefficiency, which could be prohibitive for 3D image registration.

Another solution is to progressively perform coarse-to-fine alignment via multi-scale feature maps or FP (Ranjan and Black, 2017; Sun et al., 2018; de Vos et al., 2019; Xu et al., 2021; Chang et al., 2017; Lv et al., 2019). In coarse-to-fine approach, the feature maps F_0^s and F_0^t are first aligned using a low-resolution or coarse approximation of the transformation $\phi_0(F_0^s)$. Then the transformation is progressively estimated by accumulating the residual transformation ϕ_k between the target feature map F_k^t and the warped source feature map based on previous $(k-1)$ -level registration $\phi_{k-1}(F_k^s)$ via a network \mathcal{R}_k :

$$\begin{cases} \phi_k = \phi_k \circ \phi_{k-1} \\ \phi_k = \mathcal{R}_k(\phi_{k-1}(F_k^s), F_k^t) \end{cases} \quad (4)$$

where \circ denotes the composition of two spatial transformations, and ϕ_0 is initialized as the identity transform. This approach is motivated by the fact that direct registration of original high-resolution images can be computationally expensive and prone to local optima, especially when the images are misaligned by large amounts or contain large-scale variations of motion. However, those spatial transforms from different scales are usually directly combined at each position with equal weight (Zhao et al., 2019b; Xu et al., 2021; Ranjan and Black, 2017; Chang et al., 2017). This leads to a lack of flexibility in the balance between similarity measurements of the aligned images and the anatomical rationality of the predicted motions, especially for the alignment of texture-poor areas.

Additionally, the aforementioned non-rigid registration techniques assume that the deformation field across the image is continuous and smooth. However, as shown in Fig. 2, the deformation field at the boundary between organs could be discontinuous (Hua et al., 2017), as organs can move relatively to each other in different ways, such as the motion of nearby objects (Xiao et al., 2006) or organs (Papież et al., 2014; Schmidt-Richberg et al., 2012; Vishnevskiy et al., 2016). Especially, abdomen CT scans often include large, discontinuous deformation, caused by differences in organ shapes, patient positions, and respiratory motion (Hua et al., 2017; Papież et al., 2018). This makes

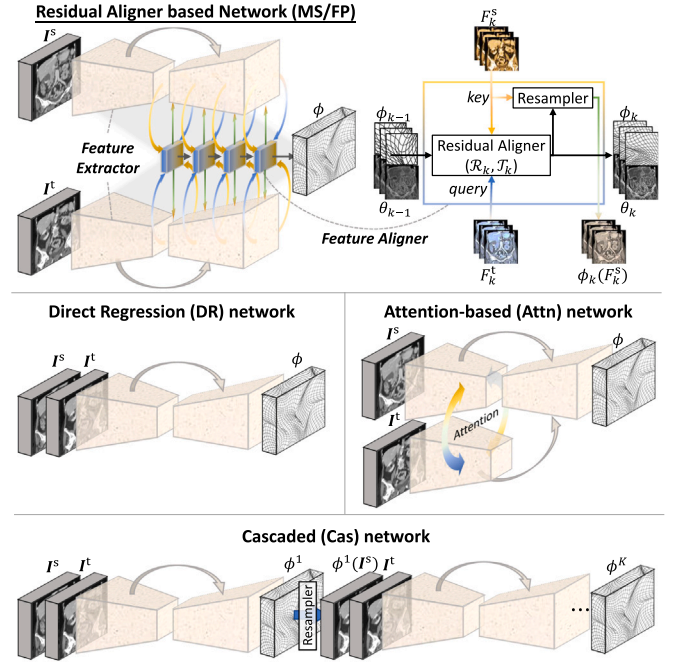


Fig. 1. The registration framework of our proposed Motion-Separable or Feature Pyramid (MS/FP) based Residual Aligner Network (RAN) structure, Direct Regression (DR), Attention-based (Attn) registration, and Cascaded (Cas) network. The MS or FP backbone networks (the difference detailed in Fig. 3) extract the features and output feature maps and the stacked RA modules (see more in Fig. 4) align and connect the data streams from the input images.

it challenging to obtain accurate registration results with current deep learning rigid or non-rigid registration methods, as these methods may fail to capture complex deformation patterns. As a result, enforcing continuity and smoothness throughout the image tends to induce errors and artifacts at the discontinuous interface, as demonstrated in Fig. 2(a).

To address this issue, previous approaches have attempted to incorporate additional information into the registration network. For example, in Cao et al. (2021), the edge map of the images is extracted using the Sobel operator and combined with the original images as input to the registration network. However, the Sobel operator can be easily approximated by a single convolution layer, leading to limited improvement. Another approach proposed in Chen et al. (2021b) is a segmentation-based discontinuous deformable registration method. It directly feeds the predicted segmentation results and incorporates them into the registration network to divide the deformation into specific sub-regions. However, this method requires additional segmentation annotation during the training phase and relies on an extra segmentation network in the inference phases, which consequently restricts its applicability to unsupervised registration. In Ng and Ebrahimi (2020), a novel regularization term was introduced to enforce constraints on the vertical projection changes between neighboring displacements while preserving the sliding motion. To the best of our knowledge, all the previous studies focus on the edge information provided by extra input information or the constraint of discontinuity in the loss function, but none of them has specifically investigated the limitation of discontinuous deformable registration in terms of network structures designed.

In this paper, we investigate the limitation in network design in the context of discontinuous deformable registration. To address these, we propose the RAN based on a novel MS backbone structure (Fig. 3) and a new RA module (Fig. 4), aiming at efficient, motion-separable, coarse-to-fine image registration. Our contributions are as follows:

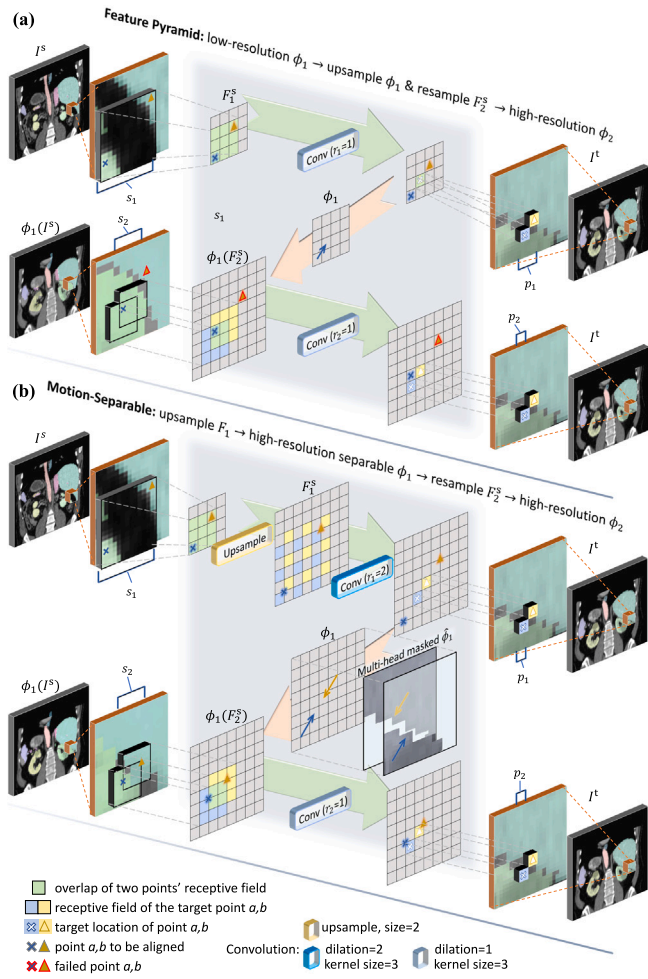


Fig. 2. Illustration of the motion inseparable problem in the coarse-to-fine alignment of two neighboring organs, with differing motions of two exemplar points (X and Δ). (a) Failed capture of point Δ is due to the low-resolution feature pyramid. (b) Our proposed solution utilizes optimized upsampling layers + dilated convolutions in the MS structure to improve the discontinuity preservation in different organs' motions (analyzed in Section 2 and detailed in Fig. 3), while maintaining the same receptive field. Additionally, a M-H mask regressed from the feature map is multiplied to disentangle the motions of different organs (detailed in Fig. 4).

- **Discontinuous deformable registration:** To the best of our knowledge, this is the first study to explore network structure design specifically for estimating the discontinuous displacement or motion field in an unsupervised manner;
- **Theoretical analysis:** The accessible displacement magnitude and the displacement discontinuity are respectively quantified and named as accessible motion capture range and motion separability pattern respectively in Sections 2.2 and 2.3, so that we provide a theoretical analysis of the upper boundary of the motion separability (Theorem 1, and Eq. (10) given in Box I), guiding the design and the parameter setting in the proposed network structure;
- **MS backbone structure:** Following Theorem 1 and Eq. (10), a new MS structure, as illustrated in Fig. 2(b), employs optimal dilated convolutions on high-resolution feature maps to benefit the network on predicting different motion patterns, while preserving the accessible motion range. Notably, this structure is computationally efficient (Section 2.4);
- **Motion disentanglement and refinement module:** Furthermore, our proposed RA module utilizes confidence and M-H mechanism based on the semantic and contextual information

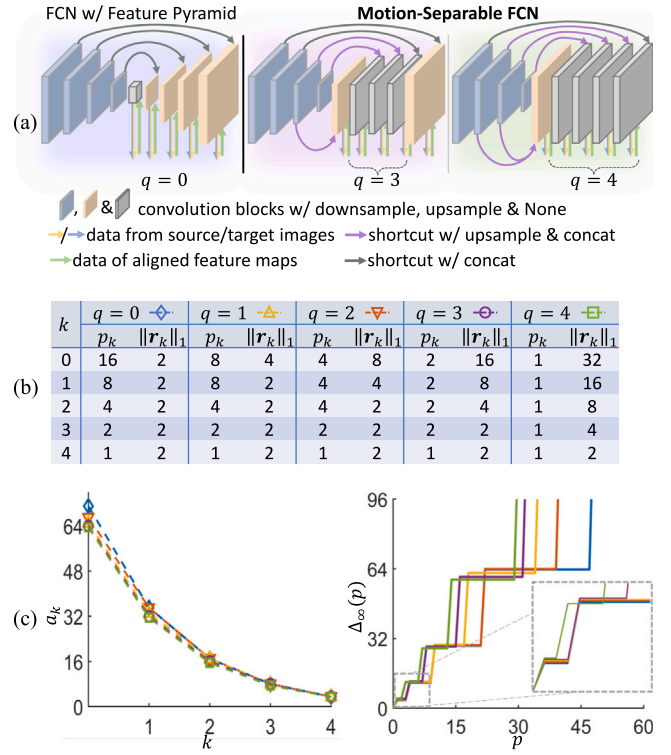


Fig. 3. The design and theoretical analysis of Fully Convolution Networks (FCN) backbone for feature extraction. (a) Motion-Separable Structures designed with a varying number of motion-separable layers q , where the feature maps from the encoder part are upsampled and concatenated to the decoder part, (b) with different hyperparameter setting, showing that a higher q , (c) with almost the same a_k , achieves a higher area under the curve of the motion separability bottleneck $\Delta_\infty(p)$, referring to Eqs. (5) and (10) (unit: pix/vox).

to disentangle the predicted displacement in different organs or regions (Section 3);

- **Accurate and Efficient registration results:** The above-proposed components constitute the novel RAN that performs efficient, coarse-to-fine, motion-separable unsupervised registration achieving state-of-the-art accuracy on publicly available lung and abdomen CT data in Section 4.

1.1. Related works

Besides the conventional iterative algorithm-based methods, the previous deep learning methods, as shown in Fig. 1, are classified into four groups, DR registration, Cascaded network (Cas) registration, Attn registration, and FP-based registration.

1.1.1. Direct regression registration methods

Voxelmorph (VM) (Balakrishnan et al., 2019) is the first deep learning method using a convolution neural network, U-net (Ronneberger et al., 2015), to directly regress the spatial transform for deformable image registration. However, the capture range of large motions by the DR-based learning methods is usually limited by the receptive field in the convolution networks between two images which thus requires downsampling layers to enlarge the receptive field or a pre-alignment.

1.1.2. Multi-stage cascaded network registration

DIRnet (de Vos et al., 2019) was thus proposed with multi-stage Cascaded (Cas) networks for coarse-to-fine registration with each network trained for the specific resolution and searching range of registration,

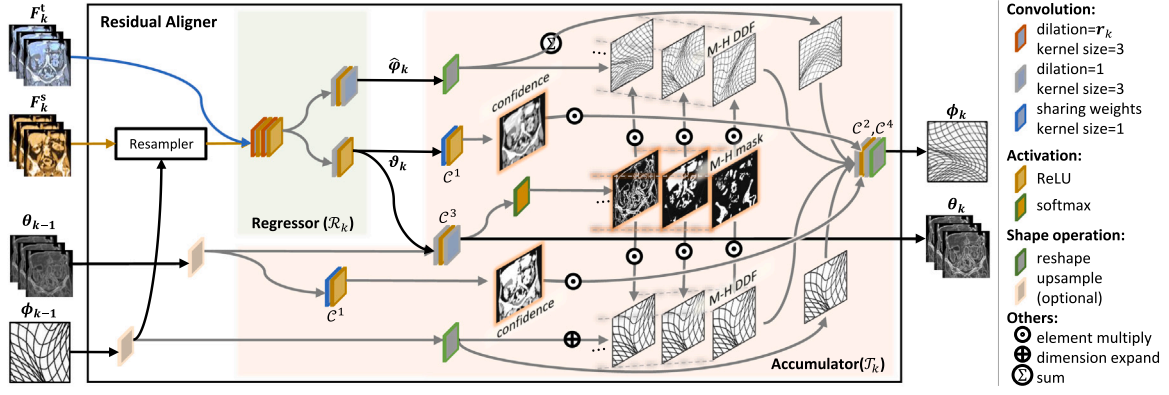


Fig. 4. The architecture of the k th Residual Aligner (RA) module. The Regressor section regresses the residual M-H DDF $\hat{\phi}_k$ and each pixel's attribute θ_k , while the Accumulator refines the DDF ϕ_k via interpolation and fusion of the M-H DDF predictions weighted by the confidence and disentangled by the M-H masks (calculated by Eq. (12)).

Table 1

Comparison between our RAN models and other models in terms of both accuracy and efficiency shows our models achieves the best performance in abdomen CT and one of the best accuracy in lung CT (both inter- and intra-subject registration), which also prove the improvement achieved by MS structure. Our models are highlighted.

Model	Reg. type	Abdomen (9 organs)				Chest (lung)				Efficiency			
		DSC↑ (%)	HD↓ (mm)	ASD↓ (mm)	detJ↓ (e3)	DSC↑ (%)	HD↓ (mm)	ASD↓ (mm)	detJ↓ (e3)	TRE↓ (mm)	#Par↓ (e6)	FLOPs↓ (e9)	TPI↓ (s)
Initial	–	30.9	49.5	16.04	–	61.9	41.6	15.86	–	10.41	–	–	–
Demons1	Iter	46.3	42.6	7.95	0.91	87.5	41.6	4.45	0.68	–	–	–	67
Demons2	Iter	49.6	41.8	7.20	0.74	90.4	32.2	3.34	0.77	–	–	–	98
VM1	•DR	44.7	43.8	9.24	2.23	84.0	32.9	6.38	5.94	3.57	0.36	34.2	0.23
VM2	•DR	51.9	45.0	8.40	4.03	88.8	32.0	5.02	15.58	2.89	1.42	69.6	0.25
CA/P	•Attn	47.6	43.8	8.77	3.85	84.7	28.9	5.75	2.67	3.18	0.58	114.5	0.41
SA/VM	•Attn	49.7	45.9	8.63	5.22	91.4	26.8	4.82	7.42	–	1.92	109.8	0.57
Cn+Un	•Cas	53.6	44.6	7.84	4.13	91.1	29.7	3.84	4.23	2.07	2.11	94.7	0.36
RCn1	•Cas	55.6	44.9	7.79	2.91	89.8	33.1	4.68	5.68	2.54	0.36	219.2	0.44
RCn2	•Cas	59.5	44.1	6.95	1.36	93.7	29.1	3.04	1.66	1.72	1.42	308.7	0.45
DPRn	•FP	53.9	57.1	8.18	4.28	88.4	29.9	4.48	3.46	2.48	0.62	82.1	0.46
RAN ₃	•MS	54.2	43.8	7.74	3.48	93.5	26.3	3.01	4.05	1.69	0.72	132.1	0.48
RAN ₄ ⁺	•MS	61.7	40.8	6.51	1.55	91.6	29.2	3.84	3.17	1.88	0.75	272.6	0.56

Table 2

Ablation study on RA module by inter-subject image registration of abdomen CT and lung CT using the baseline DPRn (Kang et al., 2022), with the varying motion-separable types of feature maps ($q = 0, 3, 4$), the M-H setting, and confidence weights (CW).

Model	Setting			Abdomen (9 organs)				Chest (lung)				Efficiency		
	CW	M-H	q	DSC↑ (%)	HD↓ (mm)	ASD↓ (mm)	detJ↓ (e3)	DSC↑ (%)	HD↓ (mm)	ASD↓ (mm)	detJ↓ (e3)	#Par↓ (e6)	FLOPs↓ (e9)	TPI↓ (s)
DPRn	×	×	0	53.4	57.1	8.18	4.28	88.4	29.9	4.48	3.46	0.62	83.2	0.46
RAN	✓	×	0	53.9	46.0	8.03	2.65	90.7	30.3	3.74	9.61	0.68	96.7	0.41
RAN	✓	×	4	56.4	44.8	7.48	2.66	92.1	28.1	3.42	6.87	0.71	201.6	0.56
RAN ₀	✓	✓	0	53.3	44.0	7.98	2.64	92.5	28.9	3.34	3.74	0.71	116.7	0.47
RAN ₃	✓	✓	3	54.2	43.8	7.74	3.48	93.5	26.3	3.01	4.05	0.72	132.1	0.48
RAN ₄	✓	✓	4	56.1	44.2	7.66	2.46	91.5	26.9	3.55	4.06	0.71	222.3	0.54
RAN ₄ ⁺	✓	✓	4	61.7	40.8	6.51	1.55	91.6	29.2	3.84	3.17	0.75	272.6	0.56

Table 3

Our unsupervised RAN₄⁺ model achieves comparable accuracy to the external segmentation-based supervised model (Chen et al., 2021b).

Model	Abdomen (9 organs)			
	DSC↑ (%)	HD↓ (mm)	ASD↓ (mm)	detJ↓ (e3)
(supervised [S] / unsupervised [U])				
Initial	30.9	49.5	16.04	–
Chen et al. (2021b) [S]	56.7	39.8	7.10	0.68
RAN ₄ ⁺ [U]	61.7	40.8	6.51	1.55

Table 4

Comparison between different regularization terms in loss function using our RAN₄⁺ model. The default setting is highlighted.

RAN ₄ ⁺ w/ loss setting	Abdomen (9 organs)			
	DSC↑ (%)	HD↓ (mm)	ASD↓ (mm)	detJ↓ (e3)
Eq. (16) w/ $\tau = 0$	56.7	42.4	7.07	1.22
Eq. (16) w/ $\tau = 25$	59.8	43.2	6.90	2.28
Eq. (15) w/ $\tau = 0$	57.8	44.6	7.17	0.94
Eq. (15) w/ $\tau = 25$	61.7	40.8	6.51	1.55

using B-spline for interpolation on the sparse prediction, but require extra time cost on training. Several end-to-end training cascaded networks (Zhao et al., 2019b; Hu et al., 2018; Shen et al., 2019; Zhao

et al., 2019a) were also proposed for coarse-to-fine image registration by recursively warping images. However, the sub-network of each stage



Fig. 5. Qualitative example in abdomen CT shows our networks achieve plausible registration, with the improvement at the areas between different organs, such as the liver, inferior vena cava, spleen, and left/right kidney. As illustrated in $|\phi|$, the discontinuous motions between different organs are preserved in the predictions by our networks. The white “ ∞ ” in the zoom-in picture of $|\phi|$ shows the motion vectors (source \mapsto target) projected in the frontal/transverse planes.

is fed with the directly warped images, which lacks feature preservation between different stages, and thus leads to extra calculation and parameters consuming on extracting the repeated features.

1.1.3. Feature pyramidal image registration

To efficiently employed the features, FP was employed for unsupervised registration in Dual-PRNet (DPRn) (Kang et al., 2022). Multiple spatial transforms are predicted in the multi-scale feature domain, to gradually refine the registration based on a sequence of feature maps extracted from a compacted structure (Kang et al., 2022). Furthermore, Edge-Aware Pyramidal Network (Cao et al., 2021) was designed for unsupervised registration with an extra edge image of the original

input to enhance the texture structure features. A new bilevel, self-tune framework (Liu et al., 2021) was also proposed for training a pyramidal-based registration network with contextual regularization. However, motion predictions in texture-free or repeatedly textured regions rely on the use of convolutional layers to infer information from neighboring information-rich regions, based on regularization terms. The inductive bias of spatial equivalence results in fixed weights for the filtering kernel, which requires extra information to differentiate between different regions. In contrast, previous methods (Cao et al., 2021; Liu et al., 2021; Kang et al., 2022) do not quantify the confidence level of different regions, making it difficult for the network to adaptively refine the prediction results to balance the similarity of the registered images with the plausibility of the predicted motion.

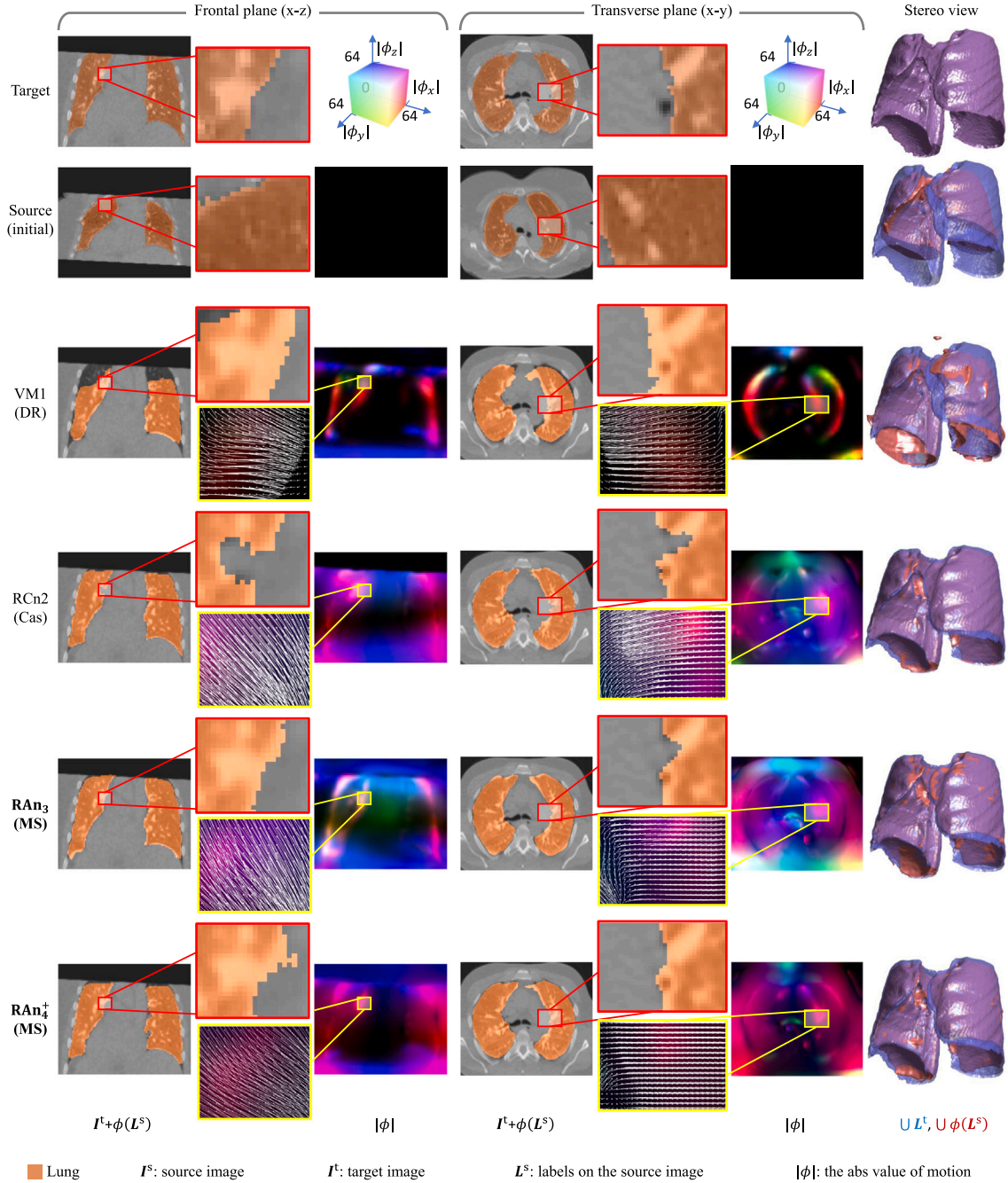


Fig. 6. Qualitative example in chest CT shows our networks achieve plausible registration, with the improvement at the edge area of the lung. The white “ \rightarrow ” in the zoom-in picture of $|\phi|$ shows the motion vectors (source \mapsto target) projected in the frontal/transverse planes.

1.1.4. Attention-based image registration

The attention mechanism (Vaswani et al., 2017) addresses the limited receptive field of CNNs and has been widely utilized in transformer networks. Optimal correspondence matching was studied in Li et al. (2021) for a stereo matching task, where a self-attention-based transformer is proposed to relax the limitation of a fixed disparity range. Local feature matching can also benefit from self- and cross-attention, because transformer networks are proven to obtain feature descriptors that are conditioned on both images (Sun et al., 2021). The attention-based mechanism was applied to deformable registration (Zhang et al., 2021; Song et al., 2021; Zheng et al., 2022; Chen et al., 2021a, 2022) previously, however, is computationally expensive for 3D image registration with $\mathcal{O}(n_p^2)$ computation complexity of n_p patches.

Based on an overview of the existing literature, there is a trade-off between computational efficiency, capturing large deformations, and preserving discontinuities between different organs’ motions. However, current existing network designs tend to prioritize the first two aspects and often neglect the importance of discontinuity preservation. Consequently, our objective is to develop an efficient network structure that addresses both discontinuity preservation and the capture of large deformations, with a particular emphasis on discontinuity preservation.

2. Network design for motion-separable structure

In this section, we first introduce the framework of coarse-to-fine registration in Section 2.1. Then the capacity of the coarse-to-fine

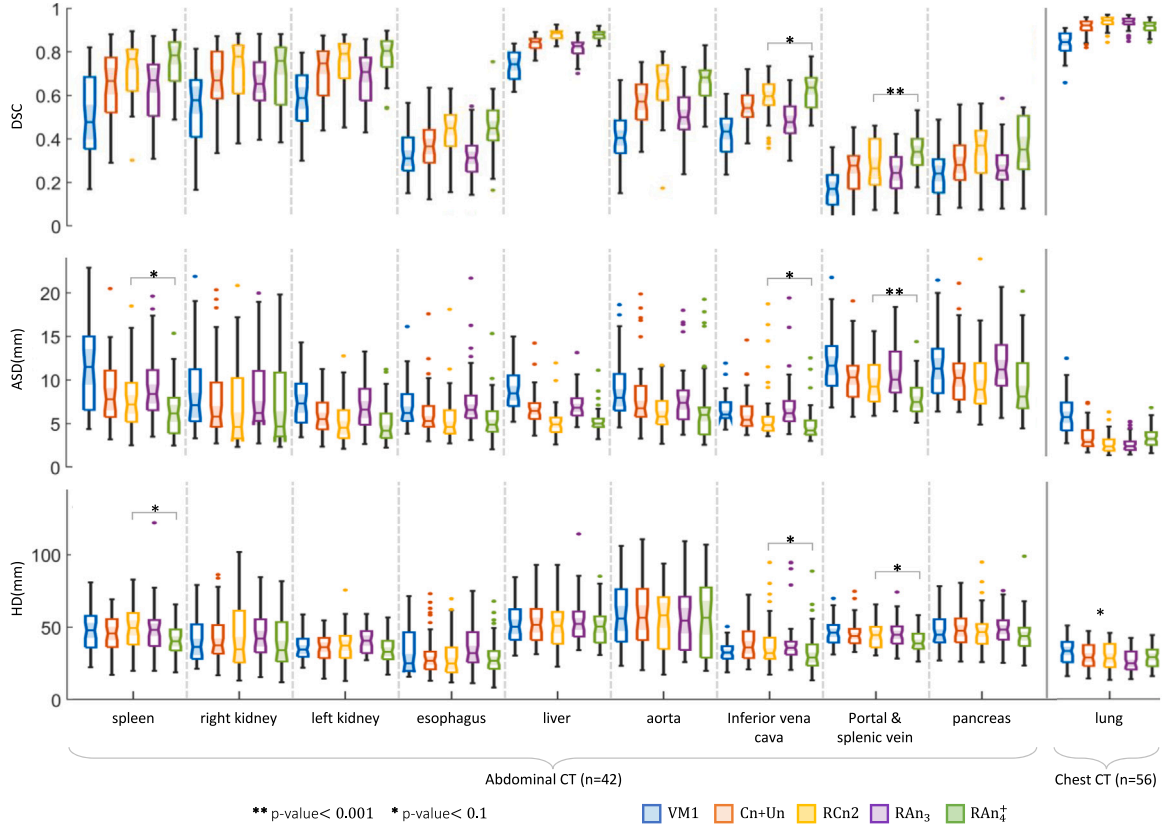


Fig. 7. Our RANs achieve the best registration in the veins in abdominal CT scans. The box plots of DSC, ASD, and HD illustrate our networks achieve the best registration in inferior vena cava and portal & splenic vein, (sample numbers 42&56 for abdomen&chest). RANs equipped with higher q perform better on the smaller organs (c.f. RAN_3 with RAN_4^+). Comparisons on more models are illustrated in [Appendix C](#).

registration networks are investigated and quantified respectively for capturing large deformation (accessible motion capture range) and preserving discontinuity (separability of the predicted motions) in Sections 2.2 and 2.3. Accessible Motion Range is defined in Section 2.2 to measure how effectively the network can handle and represent significant motion between different organs. Motion Separability, on the other hand, is defined in Section 2.3 to quantify the capacity of a registration network to maintain the distinct motion patterns of different organs and avoid blending or mixing of their motions. A bottleneck is identified in preserving discontinuity between different motions within a certain range region when using current coarse-to-fine registration.

To enhance the ability of the coarse-to-fine registration network to preserve large discontinuities and simultaneously capture large deformations, we propose a new approach based on our theoretical findings. As detailed in Section 2.4, our proposed coarse-to-fine image registration network employs a MS structure to extract feature maps. Additionally, it uses stacked progressive registration modules, or RA, to identify correspondences and estimate the DDF. These modules are further described in Section 3. The process is visualized in [Fig. 1](#).

2.1. Coarse-to-fine registration framework

As shown in [Fig. 1](#) and described by Eq. (4), the framework of coarse-to-fine registration consists of two parts, the feature extractor, and the feature aligner. A pair of Fully Convolution Networks (FCNs), with shared weights for efficient training, is used here as the feature extractor to extract two sets of feature maps, $\{F_k^s\}_{k=1}^K$ and $\{F_k^t\}_{k=1}^K$. The feature aligner, including \mathcal{R}_k and \mathcal{T}_k , takes the two feature maps from FCNs, retrieves one (key) on another (query) and then feeds back the aligned feature maps respectively to reinforce the next feature extraction.

2.2. Capture range in coarse-to-fine registration

To quantify the network's capacity for the large deformation capturing, the accessible motion capture range is defined as:

Definition 1 (Accessible Motion Range). The radius of capture range of the k^{th} -level registration by \mathcal{R}_k in Eq. (4) is defined as the smallest upper bound of its accessible DDF:

$$a_k := \min_x \{ \sup \{ \|\varphi_k[\mathbf{x}]\|_\infty \} \} \quad (5)$$

where $\|\cdot\|_\infty$ denotes the L- ∞ norm of a vector, $\sup(\cdot)$ denotes the supremum or the maximum value of a given function with varying inputs and trainable weights of networks, \mathbf{x} denotes one coordinate entry of the images or DDFs.

The accessible motion range can be approximated based on the module's receptive field: $a_k \approx \frac{s_k-1}{2}$, where s_k denotes the original-resolution size of the effective receptive field on the input feature maps, controlled by a series of dilation rates as suggested in [Zhou et al. \(2020\)](#) and the resolution:

$$s_k = \underbrace{p_k}_{(i)} \underbrace{(1 + 2\|\mathbf{r}_k\|_1)}_{(ii)} \quad (6)$$

where \mathbf{r}_k denotes the vector of dilation rates of convolution layers in the k^{th} -level registration, p_k is the minimal registration region size. This size corresponds to the pool size of the k^{th} feature map, which is determined by mapping one pixel on the original image when downsampling layers are used for spatial dimension reduction, or it corresponds to the patch size when patch-based tokenization is employed. It is required that with $p_k \leq p_{k-1}, \forall k > 0$. Additionally, the convolutions are all assumed with kernel size less than or equal to 3 to minimize the computation cost.

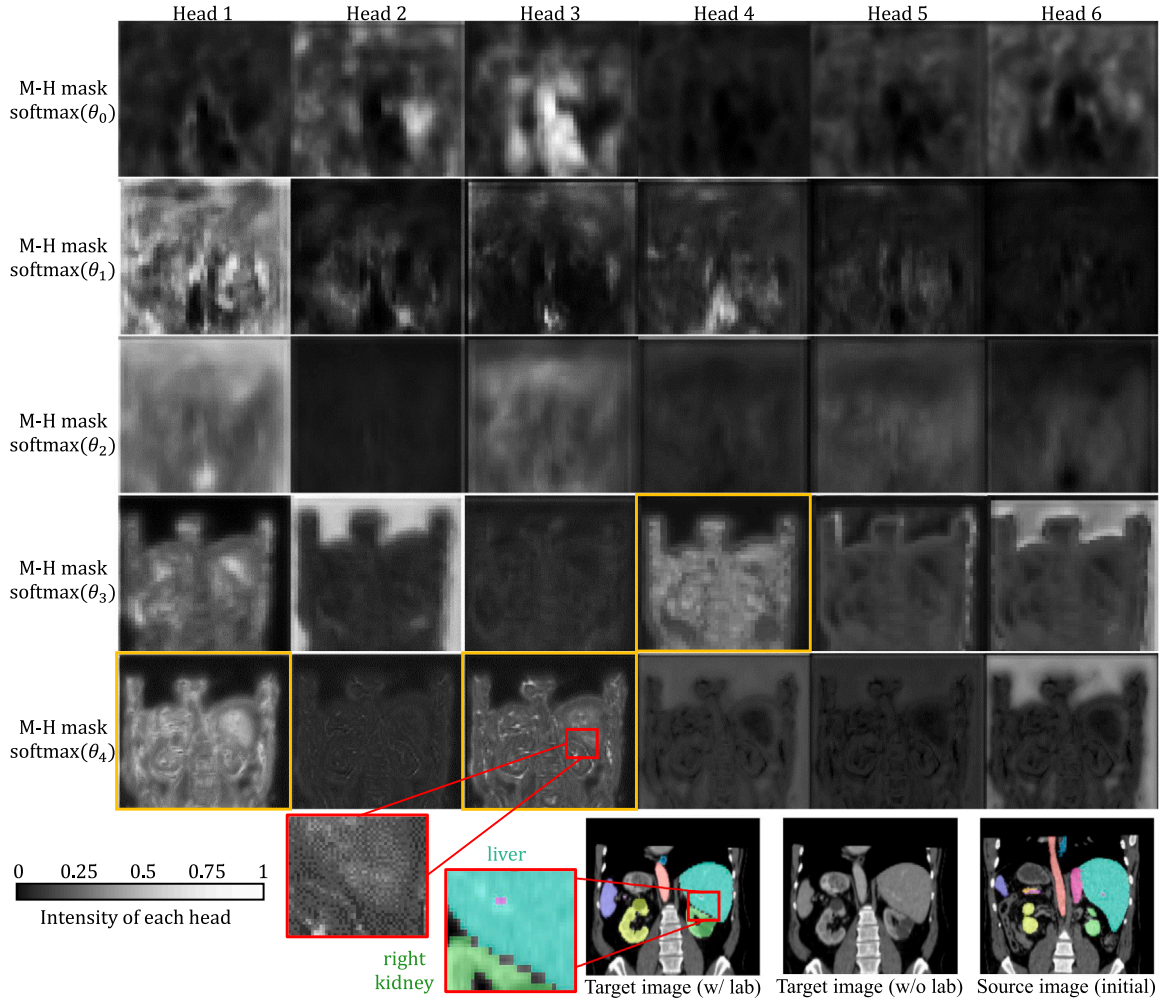


Fig. 8. Illustration of M-H Mask $\text{softmax}(\theta_k)$ at different level number k for RAN_1^+ , shows the selected regions varying from macroscopic to detail, with tissue region separated from cavity region (head-4, 3rd-level) and their edges identified (head-1&3, 4th-level).

As a result, the part (i) and (ii) in Eq. (6) are respectively dependent on pool size and dilation.

In the case of global registration on the whole image, the hyper-parameters p_1, r_1 are set to enable a_1 to reach the whole image:

$$p_1(1 + 2\|r_k\|_1) \geq 2 \max(T, H, W) + 1 \quad (7)$$

and thus accessible motion range covers the whole image.

2.3. Motion separability

In the FP approach, the typical convolution without dilation and the FP is employed: $r_k = [1 \ 1 \ \dots]$, $\forall k \in [0, K] \cap \mathbb{Z}$, $p_k = 2^{K-k}$, which fixes the Eq. (6)(ii) and relies on downsampling to enlarge receptive field with only $\mathcal{O}(n)$ complexity to reach the whole image. However, as shown in Fig. 2(a), the DDF predicted on low-resolution feature map could form a bottleneck of Degree of Freedom (DoF) of the estimated DDF. Only one predicted displacement is occupied by point \times , and thus point Δ is not retrieved until finer resolution. However, the finer stage registration focus on a smaller field of view, leading to the loss of tracking on point Δ . Here, points \times and Δ could be at the discontinuous edges of different objects or even just two tiny separate objects. To quantify the DoF limitation in the discontinuity of the estimated DDF, we define the separability of the predicted motion:

Definition 2 (Separability Bottleneck of Predicted Motion). The motion separability bottleneck is defined as the minimum value of the upper

bound of the Chebyshev difference of a network's predicted DDF ϕ between two locations $\mathbf{x}, \mathbf{y} \in \mathbb{Z}^d$ with the specific Chebyshev distance $p \in \mathbb{Z}^d$:

$$\Delta_\infty(p) := \min_{\mathbf{x}, \mathbf{y}} \{ \sup(\|\phi[\mathbf{x}] - \phi[\mathbf{y}]\|_\infty) : \|\mathbf{x} - \mathbf{y}\|_\infty = p \} \quad (8)$$

where p denotes the L- ∞ distance between the two pixels.

The reason for the problem shown in Fig. 2(a) is coarse-to-fine registration based on the FP suffering from the limited range of motion separability with respect to the capture range and the pool size.

Theorem 1 (Regional Dependency). The upper boundary of motion difference is related to a_k and p_k :

$$\begin{aligned} \forall \mathbf{x}, \mathbf{y} \in \mathbb{Z}^d, \|\mathbf{x} - \mathbf{y}\|_\infty \geq p_{k''} + 2 \sum_{k'=k''+1}^k a_{k'}, \\ \sup(\|\phi_k[\mathbf{x}] - \phi_k[\mathbf{y}]\|_\infty) \geq 2 \sum_{k'=k''}^k a_{k'}; \\ \exists \mathbf{x}, \mathbf{y} \in \mathbb{Z}^d, \|\mathbf{x} - \mathbf{y}\|_\infty < p_{k''-1} + 2 \sum_{k'=k''}^k a_{k'}, \\ \sup(\|\phi_k[\mathbf{x}] - \phi_k[\mathbf{y}]\|_\infty) = 2 \sum_{k'=k''}^k a_{k'}; \end{aligned} \quad (9)$$

where k'', k , denote two recursive numbers satisfying $0 \leq k'' < k$, and \mathbf{x}, \mathbf{y} denote two coordinate entries of images/DDFs.

Following Theorem 1, $\Delta_\infty(p)$ is calculated as in Eq. (10) to describe the limitation on the multi-objects' motion difference. We substitute the Eq. (6) into Eq. (10) and thus can easily derive the relation between motion separability bottleneck Δ_∞ and the pool sizes (p_1, \dots, p_K) as well as the dilation rates (r_1, \dots, r_K). The proof of Theorem 1 is given in Appendix A.1.

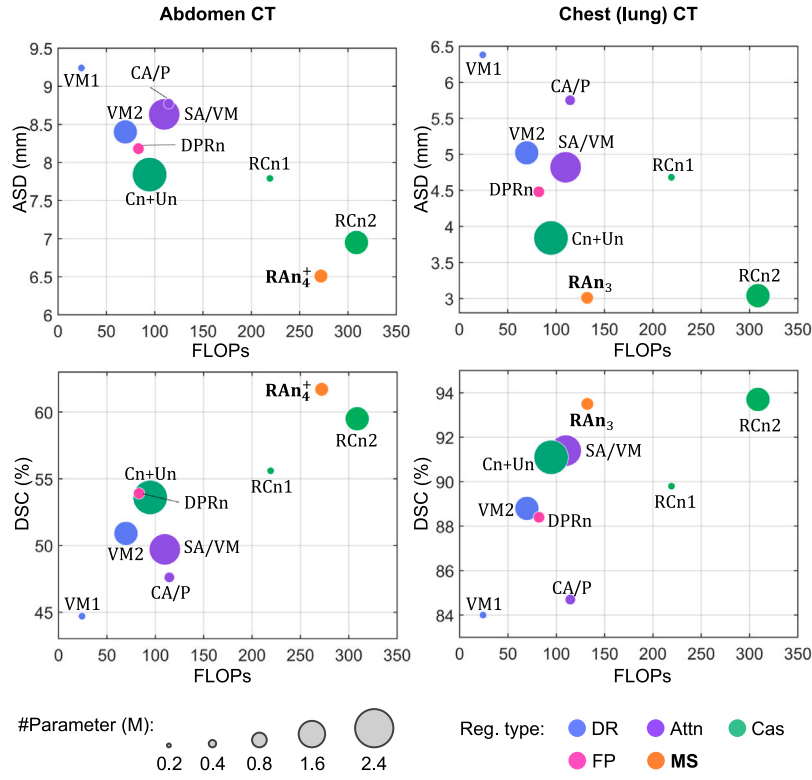


Fig. 9. Our RAn₄⁺ and RAn₃ respectively achieve the best accurate registration in abdominal 9-organ CT and one of the best accuracy in lung CT, with better efficiencies (the model sizes are represented by the circle size).

$$\Delta_{\infty}(p) = \begin{cases} 2 \sum_{k=1}^K a_k, & p \geq p_1 + 2 \sum_{k=2}^K a_k \\ 2 \sum_{k'=k}^K a_{k'}, & p_k + 2 \sum_{k'=k+1}^K a_{k'} \leq p < p_{k-1} + 2 \sum_{k'=k}^K a_{k'}, \quad 1 < k \leq K \\ 0, & p < p_K \end{cases} \quad (10)$$

Box I.

2.4. Motion-separable structure

According to Theorem 1, the smaller pool size releases a higher range of motion difference. Thus, we design a new backbone structure, called MS FCNs, to achieve a high DoF of DDF but still with the same capture range using dilation convolution on upsampled feature maps as shown in Fig. 3(a). Different from the previously used FP-based FCNs, the shortcut feature maps from the encoder part are upsampled and concatenated to a specific high-resolution feature map as the input to the decoder part with $p_k = 2^{K-q}$, $\forall k \leq q$ and $p_k = 2^{K-k}$, $\forall q < k \leq K$, where q denotes the layer number with MS pattern in the decoder part. The q could be adjusted considering the balance between the DoF of the predicted DDF and computational cost. The complexity required is $\mathcal{O}(n \log(n))$ using fully MS-layer structure $q = K$ and is still $\mathcal{O}(n)$ using fully FP $q = 0$. To keep the same receptive field of MS structure as FP structure, the dilation rate is set to $\|r_k(q > 0)\|_1 \geq 2^{q-k} \|r_k(q = 0)\|_1$, $\forall k \leq q$ as suggested by Eqs. (5) and (6). As shown in Fig. 2(b), with the same receptive field, the MS structure releases the higher resolution before alignment and thus avoids loss of the DoF of DDF. The capture ranges and the motion separability of DDF for varying settings are illustrated in Fig. 3(b)(c)(d) based on the calculation of Eqs. (5) and (10), where the new design achieves a larger area under $\Delta_{\infty}(p)$ with almost the same a_k .

The detailed architecture of the encoder and the decoder part in our proposed networks used for the following experiments are shown in Tables B.5 and B.6.

3. Residual aligner module

The RA module as shown in Fig. 4 aims to establish spatial transform ϕ between two images via recursively warping feature map of one towards the others, based on Eq. (4). Extra attributes map θ is estimated by Regressor network \mathcal{R} , restoring the auxiliary information, to disentangle and fuse the alignment of pixels from the different anatomic regions via the Accumulator network \mathcal{T} , and thus improve the accuracy of the discontinuous motions:

$$\begin{cases} (\phi_k, \theta_k) = \mathcal{T}_k(\hat{\phi}_k, \vartheta_k, \phi_{k-1}, \theta_{k-1}) \\ (\hat{\phi}_k, \vartheta_k) = \mathcal{R}_k(\phi_{k-1}(F_k^s), F_k^t) \end{cases} \quad (11)$$

for the RA modules cascade number $k = 1, \dots, K$, where $\hat{\phi}_k$ denotes the M-H residual DDF. The k th RA module first takes the input feature maps from source images and target images $F_k^s, F_k^t \in \mathbb{R}^{c_k \times n_k}$ and use the Regressor \mathcal{R}_k to regress a m -head residual DDF $\hat{\phi}_k \in \mathbb{R}^{d_m \times n_k}$ and the incremental attributes $\vartheta_k \in \mathbb{R}^{m \times n_k}$ (Section 3.1). Then the Accumulator Network \mathcal{T} computes the confidence and M-H masks, describing the prediction's reliability and the semantic properties of each pixel, and fuse the m -head DDF weighted based on the attribute maps $\theta_k \in \mathbb{R}^{m \times n_k}$ as in Section 3.2. The warping function performed by the resampler is implemented following the work (Jaderberg et al., 2015).

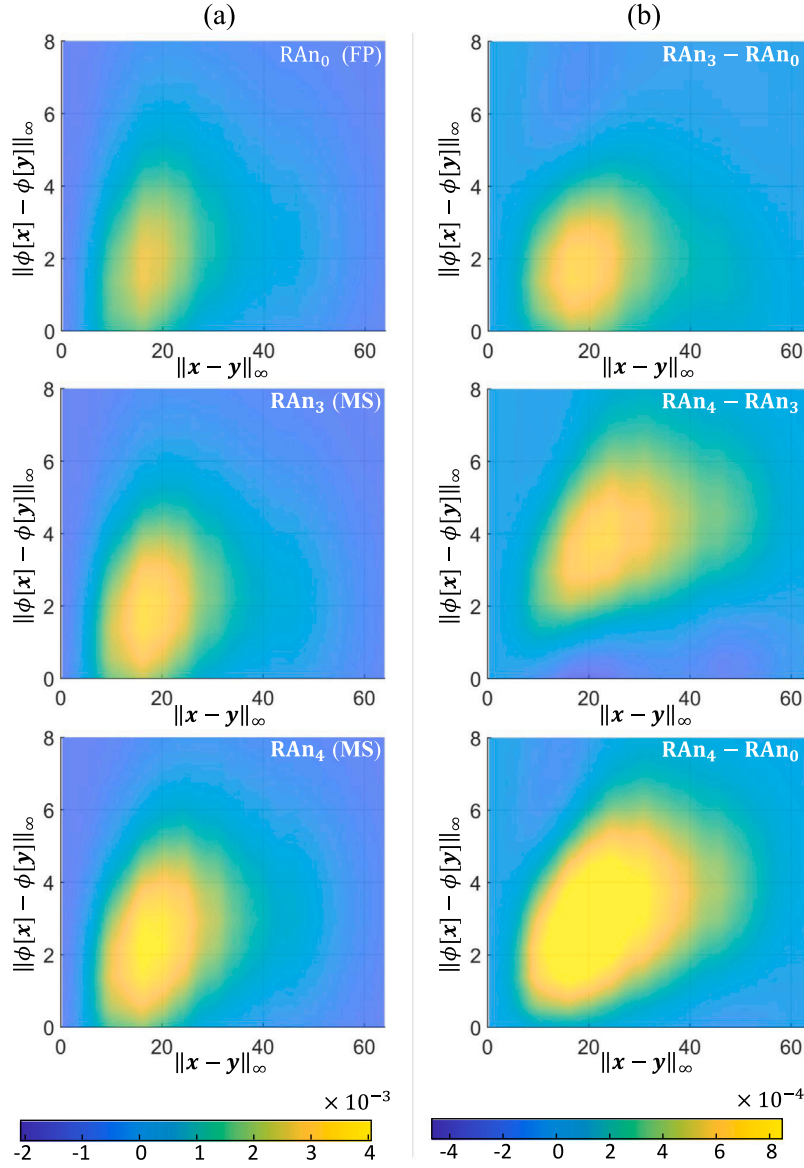


Fig. 10. (a) Empirical Probability Density Function (PDF) of the pairs of correct predicted motions by RAn_0 , RAn_3 and RAn_4 , smoothed by Gaussian filter ($\sigma = 1$ pix), with respect to varying Chebyshev distance ($\|x - y\|_\infty, \forall x, y \in \{x | L^1[x] = \phi(L^s)[x]\}$) and varying Chebyshev distance between their motions ($\|\phi[x] - \phi[y]\|_\infty$), and (b) the difference between each pair of PDF, validating that higher number of MS pattern layers enable the network to achieve better motion separability with similar model scale, where $L^{\text{s&t}} \in \{\text{spleen}, \dots, \text{pancreas}\}^n$ denote the labels on source&target images of abdomen CT.

3.1. Regressor

The function of the Regressor \mathcal{R}_k in RAN is to regress the M-H residual transform $\hat{\varphi}_k$ between the target feature map F_k^t and the source feature map warped by the previous alignment $\phi_{k-1}(F_k^s)$, with the incremental attribute map ϑ_k to restore the auxiliary information for the inter-scale refinement in the coarse-to-fine registration. As shown in Fig. 4, Regressor concatenates the input feature maps and feeds them into the subsequent series of dilated convolution and activation layers. Referring to Section 2.4, the dilation rate vector r_k is set to enlarge the capture range of alignment and raise the feature resolution as introduced in Section 2. Then two shallow convolution networks are respectively used to predict the M-H DDF and the incremental attributes raised from this level's alignment.

3.2. Accumulator

The task of Accumulator \mathcal{T}_k is to refine the DDF with the previous coarse DDF by disentangling, interpolating, and fusing those spatial

transform representations from varying scales and different heads in terms of the contextual information, such as the alignment reliability of the neighboring pixels and their semantic attributes. The calculation of Accumulator shown in Fig. 4 can be written as:

$$\begin{cases} \phi_k = C^4([\varphi'_k, \sum_{\{m\}}(\hat{\varphi}_k), \phi'_{k-1}, \phi_{k-1}]) \\ \theta_k = C^3([\vartheta_k, \theta_{k-1}]) \end{cases} \quad (12)$$

where ϕ_{k-1} and φ_k are the weighted DDF and residual DDF:

$$\begin{cases} \phi'_{k-1} = C^2(\phi_{k-1} \otimes \underbrace{\text{softmax}(\theta_k)}_{\text{M-H mask}}) \odot \underbrace{C^1(\theta_{k-1})}_{\text{confidence}} \\ \varphi'_k = C^2(\hat{\varphi}_k \odot \underbrace{\text{softmax}(\theta_k)}_{\text{M-H mask}}) \odot \underbrace{C^1(\vartheta_k)}_{\text{confidence}} \end{cases} \quad (13)$$

$\sum_{\{m\}} : \mathbb{R}^{d \times m \times n_k} \rightarrow \mathbb{R}^{d \times n_k}$ denotes the head-dimension sum, $\otimes : \mathbb{R}^{d \times n_k} \times \mathbb{R}^{m \times n_k} \rightarrow \mathbb{R}^{d \times m \times n_k}$ denotes the tensor product, $\odot : \mathbb{R}^{d \times m \times n_k} \times \mathbb{R}^{1 \times n_k} \rightarrow \mathbb{R}^{d \times m \times n_k}$ denotes the element-wise product for the last two dimensions. Here C^1, C^2, C^3, C^4 are fitted by convolution networks with activation

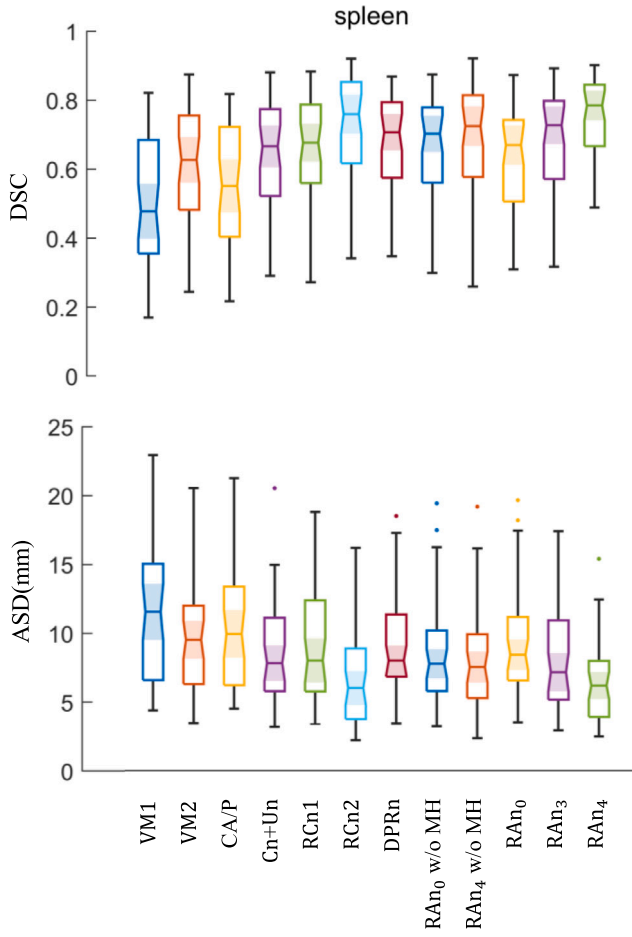


Fig. B.11. Results on spleen.

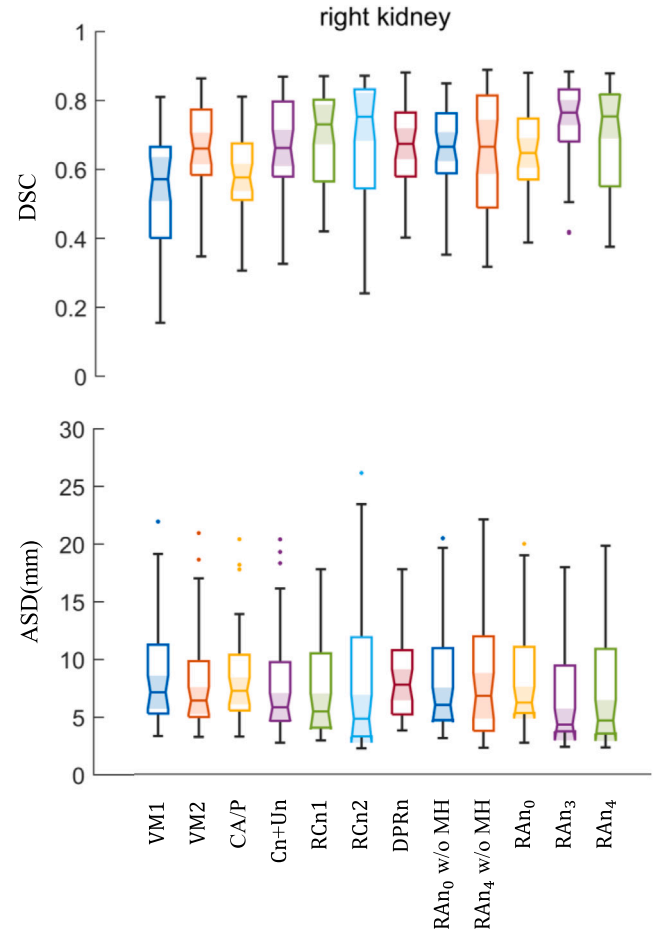


Fig. B.12. Results on right kidney.

layers, respectively for the mapping of confidence weight projection, interpolation, attribute fusion, and the DDF fusion.

3.2.1. Confidence weight for motion refinement

Simple composition of DDFs from different levels (de Vos et al., 2019; Xu et al., 2021) could accumulate errors at the points which failed in the previous alignment. Thus, the confidence values are respectively quantified by $C^1(\vartheta_k), C^1(\vartheta_{k-1})$ in Eq. (13) for M-H residual DDF $\hat{\phi}_k$ and previous level DDF ϕ_{k-1} to adaptively weight the following filtering, performed by C^2 and C^4 , for smoothing or interpolation with neighboring prediction value. Here the confidence is implicitly regressed from ϑ_k and ϑ_{k-1} (contrary to the confidence of occlusion probability in Li et al. (2021)) with general representation aiming to provide higher accuracy.

3.2.2. Motion disentanglement via multi-head masks

Inspired by the M-H attention (Vaswani et al., 2017), the corresponding M-H masks are regressed by softmax(θ_k) to implicitly disentangle the prediction of multiple objects with different motion patterns from the M-H residual DDFs, and thus decouple the filtering on the different objects as shown in Eq. (13). This process could be regarded as the combination of DDFs selected by the M-H masks, with preserving discontinuities in the DDF and the trend of motions (Heinrich et al., 2013).

More explanation detail of confidence-weighted and M-H disentangled filtering can be found in Appendices A.2 and A.3.

4. Experiments

4.1. Datasets

We evaluated the RAN on unsupervised deformable registration on two publicly available datasets with segmentation annotations on 9 small organs in abdomen CT and lung CT:

4.1.1. Unpaired abdomen CT

The dataset is provided by Hering et al. (2021). The ground truth segmentations of spleen, right kidney, left kidney, esophagus, liver, aorta, inferior vena cava, portal, splenic vein, pancreas of all scans are provided. The deformable registration of abdominal CT imaging is considered challenging due to the large relative motion variations across disjunct tissues and the great variability in organ volume, ranging from 10 milliliters (esophagus) to 1.6 liters (liver). All scans were captured during portal venous contrast phase with variable volume sizes ($512 \times 512 \times 53 - 512 \times 512 \times 368$) and field of views (approx. $280 \times 280 \times 225 \text{ mm}^3$ to $500 \times 500 \times 760 \text{ mm}^3$). The in-plane resolution varies from $0.54 \times 0.54 \text{ mm}^2$ to $0.98 \times 0.98 \text{ mm}^2$, while the slice thickness ranged between 1.5 mm – 7.0 mm. Each volume was resized to $2 \times 2 \times 2 \text{ mm}^3$ in the pre-processing. From totally 30 subjects, 23 and 7 are respectively used for training and testing, for 506 and 42 different pairing cases, following the default setting in Fu et al. (2020).

4.1.2. Unpaired/paired chest (lung) CT

The dataset is provided by Hering et al. (2020, 2021). The CT scans are all acquired at the same time point of the breathing cycle and here

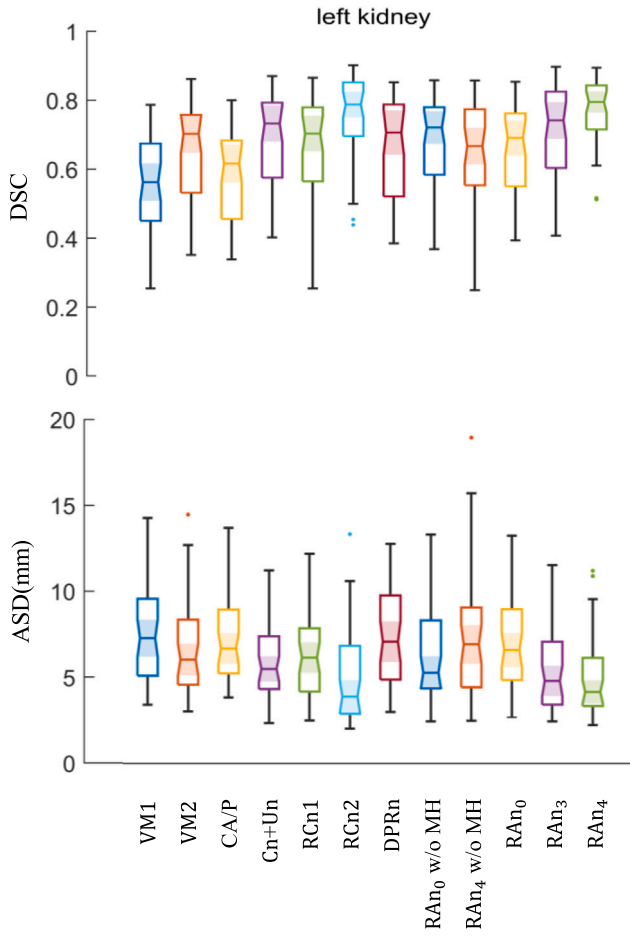


Fig. B.13. Results on left kidney.

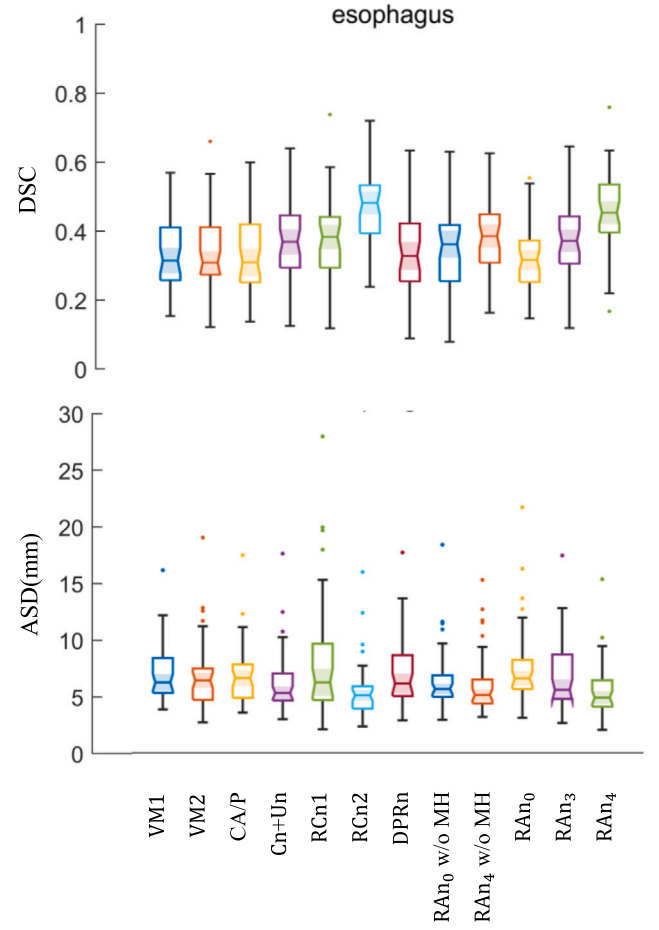


Fig. B.14. Results on left kidney.

we perform inter-subject (exhale) and intra-subject (exhale↔inhale) registration. The ground truth lung segmentation of all scans are provided. Each volume was resized to $1.75 \times 1.75 \times 1.75 \text{ mm}^3$ in the pre-processing. From the total of 20 subjects, 12 and 8 are respectively used for training and testing in intra-subject registration, and thus 132 and 56 different pairing cases in inter-subject registration, following the default setting in [Fu et al. \(2020\)](#).

4.2. Training details

We normalize the input image within 0–1 range and augment the training data by randomly cropping input images during training.

4.2.1. Synthetic training

All the models are first pre-trained for 50k iteration on synthetic DDF $\tilde{\phi}$ combining rigid spatial transformation with rotation angle $\beta \sim \mathcal{U}(-\pi/4, \pi/4)$ at an arbitrary axis and deformation synthesized by thin plate spline as well as Gaussian deformation by 20 random seeds located uniformly randomized within the image domain. , with the loss function set as:

$$\mathcal{L}_{\text{syn}} = \sum_{\mathbf{x}} \|\phi[\mathbf{x}] - \tilde{\phi}[\mathbf{x}]\|_2^2 + \lambda \sum_{\mathbf{x}} \|\nabla \phi[\mathbf{x}]\|_2^2 \quad (14)$$

where λ denotes the weight of regularization, $\mathbf{x}, \mathbf{y} \in \mathbb{Z}^d$ denotes the entries coordinates for images or DDF.

4.2.2. Real-data training

Then all the inter-subject registration models are trained on real data for 100k iterations with the same loss function which will be shown later. The final trained models are selected as the used model, since the loss value for those models all converged before 100k iterations.

Inter-subject Registration The loss function used for inter-subject registration is as following:

$$\mathcal{L}_{\text{inter}} = \underbrace{D(I^t, \phi(I^s))}_{\text{dissimilarity term}} + \lambda \underbrace{\sum_{\mathbf{x}} \|\nabla \phi[\mathbf{x}]\|_2^2 e^{-\tau \|\nabla I^t[\mathbf{x}]\|_2^2}}_{\text{regularization term}} \quad (15)$$

where local normalized cross correlation and mean squared error are used in abdomen and lung CT respectively for D following [Balakrishnan et al. \(2019\)](#). The smoothing regularization term includes an additional edge-aware factor, which is adopted from a stereo matching method ([Heise et al., 2013](#)), and τ is the temperature scaling value to adjust the intensity of smoothing regularization at the edges, which is the same L2-norm smoothing term as in [Balakrishnan et al. \(2018\)](#) when $\tau = 0$.

In addition to the edge-aware smoothness regularization term, we also tried the sliding-preserving discontinuous regularization term proposed in ([Ng and Ebrahimi, 2020](#)):

$$\mathcal{L}_{\text{inter-sp}} = \mathcal{L}_{\text{inter}} + \underbrace{\frac{\lambda_1}{2} \sum_{\mathbf{x}, \mathbf{y}} \log(1 + \|\phi[\mathbf{x}] \wedge \phi[\mathbf{y}]\|_2^2) \mathcal{K}_4(\mathbf{x}, \mathbf{y})}_{\text{sliding-preserving term}} \quad (16)$$

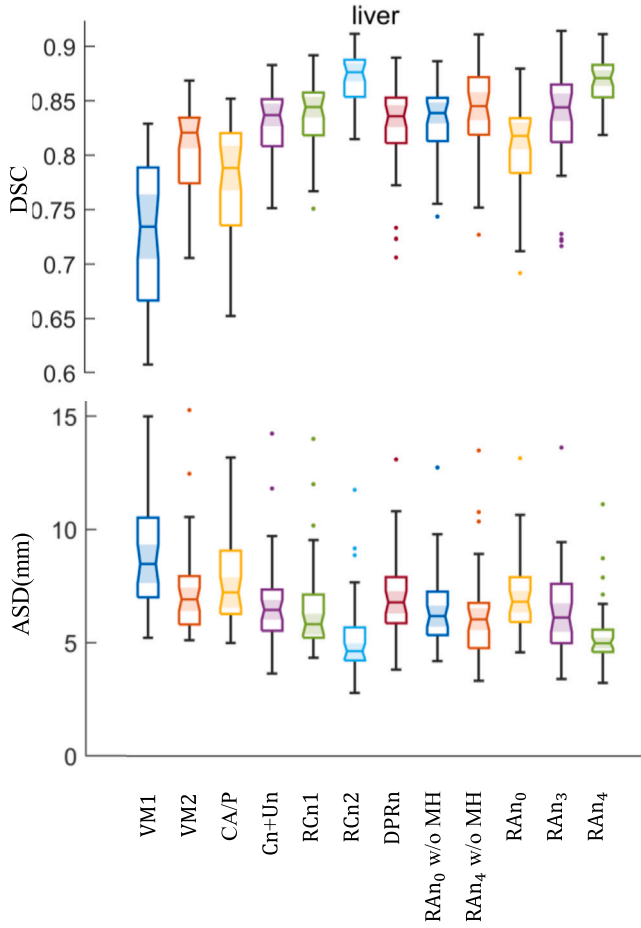


Fig. B.15. Results on liver.

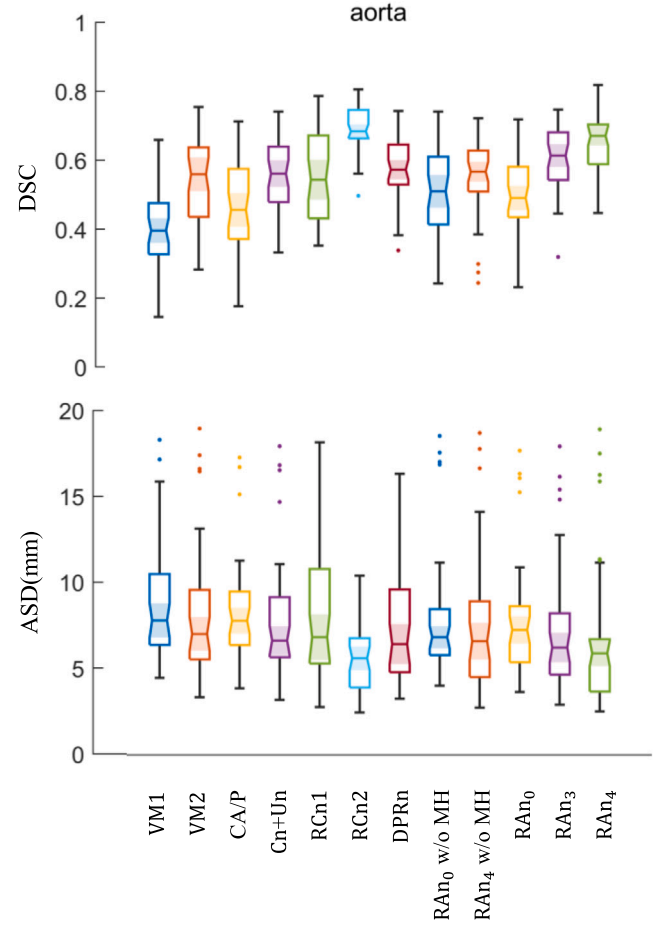


Fig. B.16. Results on aorta.

where \wedge denotes the cross product, and \mathcal{K} denotes the C^4 Wendland kernel (Wendland, 1995), and this loss setting is the equivalent to that in Ng and Ebrahimi (2020) when $\tau = 0$.

Intra-subject Registration (Lung) The loss function for training of intra-subject registration models includes one more landmark error term than Eq. (15):

$$\mathcal{L}_{\text{intra}} = \mathcal{L}_{\text{inter}} + \beta \underbrace{\frac{1}{|X|} \sum_{(x,y) \in X} \|(x-y) - \phi[y]\|_2^2}_{\text{landmark error}} \quad (17)$$

where X denotes the set of the corresponding landmarks' coordinates from the pairs of lung CT scans. We believe it is more appropriate to refer to this regularization term as “sliding-preserving” instead of “discontinuity-preserving” as claimed in Ng and Ebrahimi (2020).

4.3. Implementation and evaluation

4.3.1. Implementation

The code for image registration tasks were developed based on the framework of Balakrishnan et al. (2018) in Python using Tensorflow and Keras. It has been run on Nvidia Tesla P100-SXM2 GPU with 16 GB memory, and Intel(R) Xeon(R) Gold 6126 CPU @ 2.60 GHz. The backbone FP network we used is U-net (Ronneberger et al., 2015) based on residual structure (He et al., 2016) with four downsampling blocks and four upsampling blocks. Since the most motion difference ranges between 0–15 as shown in Fig. 10, two models RAN_3 and RAN_4^+ are thus selected as our representative models with $q = 3, 4$ as suggested by the effect shown in Fig. 3(d). The details of those structures are described in Appendix B.

4.3.2. Comparison

We compared RANs with the relevant state-of-the-art network structures in Table 1. Original Demons (Thirion, 1998) (Demons1) and the improved Demons based on fast symmetric forces (Vercauteren et al., 2009) (Demons2) implemented by Vercauteren et al. (2008) are utilized as the representative traditional iterative (Iter) registration methods. The Voxelmorph (Balakrishnan et al., 2019) (VM1/VM2: light-/heavy-weight model) is adopted as the representative method of DR. The composite network combining CNN (Cn: Global-net) and U-net (Un: Local-net) following to Hu et al. (2018), as well as 5-Recursive Cascaded network (RCn) (Zhao et al., 2019a) (RCn1/RCn2: light-/heavy-weight model) are also adopted into the framework as the relevant baselines representing multi-stage cascaded (Cas) networks. DPRn (Kang et al., 2022) is selected as the baseline for FP networks. Additionally, we also replace RA-module (in Fig. 1) with cross attention (Attn) (Vaswani et al., 2017) to compare the performance at module-level. We also incorporate the self-attention mechanism on VM2 (Balakrishnan et al., 2019) following Chen et al. (2021a) to compare the improvement from attention mechanism with MS structure.

Another supervised deformable registration method based on external segmentation is compared in Table 3.

To provide a more comprehensive evaluation, we have also included varying settings of loss function Eq. (15), and Eq. (16) with extra regularization term from Ng and Ebrahimi (2020) in our experimental settings, which is shown in Table 4.

4.3.3. Evaluation metrics

Following de Vos et al. (2019), we calculate the DSC, HD, and ASD on the annotated masks for the inter-subject registration evaluation

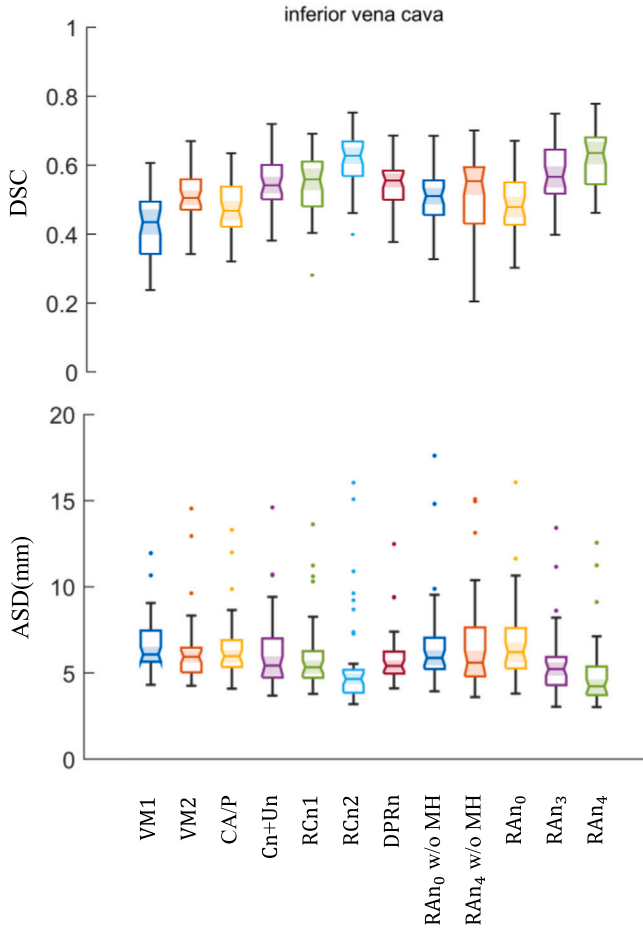


Fig. B.17. Results on inferior vena cava.

of nine organs in abdomen CT and one organ (lung) in chest CT, with the negative number of Jacobian determinant in tissues' region ($\det J$) for rationality evaluation on prediction. Target Registration Error (TRE/mm) is used for evaluation of intra-subject registration for paired lung CTs (exhale and inhale). In addition, the average values of Time cost Per Image (TPI), Parameter Number (#Par), and Float Operations (FLOPs) are used to evaluate the models' efficiency.

4.4. Results

The comparison between RANs (RAN_3 , RAN_4^+) with other methods on abdomen and lung CT using all 10 organs is shown in Table 1, which is visualized in Fig. 9. Table 2 demonstrate the quantitative results for the ablation studies. Figs. 5 and 6 respectively illustrate several examples comparing our methods with others. The separate evaluation of the 9 + 1 organs (abdomen+lung) for five models, as shown in Fig. 7

Additional registration results for specific organs performed by each model are shown in Appendix C.

4.4.1. Registration network comparison

The comparison between RANs with other unsupervised registration methods on abdomen and lung CT using all 10 organs is shown in Table 1, and the results illustrate our network achieved one of the best performances in this task with fewer parameters and lower computational cost, which is also visualized in Fig. 9. Table 3 reveals that our RAN even attains accuracy comparable to that of the external segmentation-based supervised method.

Figs. 5 and 6 highlight the superior performance of our RAN, especially in the area containing multi-organs and at the edges of organs.

It is worth noting that the discontinuities between different organs are preserved in the predictions ϕ by our RANs as shown in Fig. 5. They verify the effectiveness of our method to solve the motion separability limitation problem as previously analyzed in Section 2.3

Considering different registration types, DR networks (e.g., VM1, VM2) necessitate more parameters to achieve satisfactory results. In contrast, multi-stage Cascaded networks either demand greater computation (e.g., RCn1, RCn2) or more parameters (e.g., Cn+Un). The FP-based network (DPRn) strikes a balance between these requirements, and our MS-based RANs further refine this balance, as illustrated in Fig. 9.

A distinct evaluation of the 9 + 1 organs (abdomen+lung) for five models, showcased in Fig. 7, reveals that our RAN offers the highest accuracy in registering smaller organs (like veins) and ranks among the best in registering other organs. Moreover, when compared to the additional segmentation-based supervised method in Table 3, our RAN maintains comparable accuracy on abdominal CT scans.

4.4.2. Ablation study

To validate the effect of each component on the performance, we also tried several combinations on the confidence weight (CW), M-H and MS pattern number (q) on experiments of abdomen and lung CT as shown in Table 2. For a fair comparison, the channel numbers are tuned to keep the trainable parameter numbers similar to each others, except RAN_4^+ with larger model size for higher accuracy. Figs. 5–7 show our RAN_4^+ with $q = 4$ is better than RAN_3 on smaller tissues' prediction but worse on larger one's (lung).

4.4.3. Separability of the predicted motions

Besides the improvement from higher- q MS structure implicitly validated by the better results, more visual validation of MS design is illustrated in Fig. 10(a) including the probability density distributions of the pairs of correct motion prediction with varying voxel distance and motion difference for varying q . Based on the difference between them in Fig. 10(b), It shows RANs with higher q obtain more correct alignment hits at the left-top area, and thus the better motion separability, matching the expectation illustrated in Fig. 3(d) and validating the improvement by the design principle described in Section 2.4.

4.4.4. Multi-head mask

The M-H Mask maps at different registration levels for one example are illustrated in Fig. 8, showcasing the $\text{softmax}(\theta_k)$ maps at various level numbers k for RAN_4^+ . These maps illustrate the selected regions indicated by the masks, ranging from large-scale to small-scale and capturing both global and local features. With reference to the target image and labels, it can be observed that the head-3 mask at the 3rd-level effectively separates tissue regions from cavity regions. Furthermore, the head-1 and head-2 masks at the 4th-level further identify the tissue edges. These results demonstrate that the M-H mask successfully identifies different tissue regions, disentangles the motion by these regions, and fulfills the requirement of coarse-to-fine registration.

5. Discussion and conclusion

The limitations of the coarse-to-fine deformable image registration were studied. We found that the separability of motions is limited, and we provided a quantitative analysis of the upper bound of the separability as shown in Theorem 1 and Eq. (10). To address this limitation, a novel RAN design was introduced, which leverages a new MS structure to increase the motion separability and new RA modules to disentangle and refine the predicted DDF across different organs/regions.

The results presented in Fig. 7 show that the RANs achieved the best registration accuracy for small organs, e.g. veins, in abdominal CT scans and comparable registrations with other state-of-the-art networks for other organs in abdominal and lung CT scans, with fewer parameters

Table B.5

Network of the encoder. (ker: kernel size, dila: dilation rate).

Layer(s)	ker	dila	channels	scale	in	out
conv,norm,act	3	1	1/8	1	$I^{s,1}$	r1
conv,norm,act	3	1	8/8	1	r1	f1
conv,norm	3	3	8/8	1	f1	f1
act	–	–	8/8	1	f1+r1	s1
downsample	–	–	–	–	s1	s1
conv,norm,act	3	1	8/16	2	s1	r2
conv,norm,act	3	1	16/16	2	r2	f2
conv,norm	3	3	16/16	2	f2	f2
act	–	–	16/16	2	f2+r2	s2
downsample	–	–	–	–	s2	s2
conv,norm,act	3	1	16/16	4	s2	r3
conv,norm,act	3	1	16/16	4	r3	f3
conv,norm	3	3	16/16	4	f3	f3
act	–	–	16/16	4	f3+r3	s3
downsample	–	–	–	–	s3	s3
conv,norm,act	3	1	16/32	8	s3	r4
conv,norm,act	3	1	32/32	8	r4	f4
conv,norm	3	3	32/32	8	f4	f4
act	–	–	32/32	8	f4+r4	s4
downsample	–	–	–	–	s4	s4

and less computation. The impact of each component was validated in Table 2, including the MS structure with varying q , the motion disentanglement based on the M-H mask, and the DDF refinement based on confidence weights.

Compared with other different registration structure types in Fig. 9, the MS structure achieves the best trade-off between the registration accuracy, the computation scale, and the model size.

Additionally, the proposed RANs based on MS structure demonstrated improved separability of predicted discontinuous motion, as shown in Figs. 5 and 10 further indicated that the larger MS pattern number, q , leads to the better motion separability. These findings support the previously illustrated design principle on MS mechanism in Fig. 2 and the theoretical analysis of motion separability in Fig. 3.

These results demonstrate the efficiency and the potential of RANs performing relevant tasks including multi-object registration, which could also be further applied to other relevant tasks.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

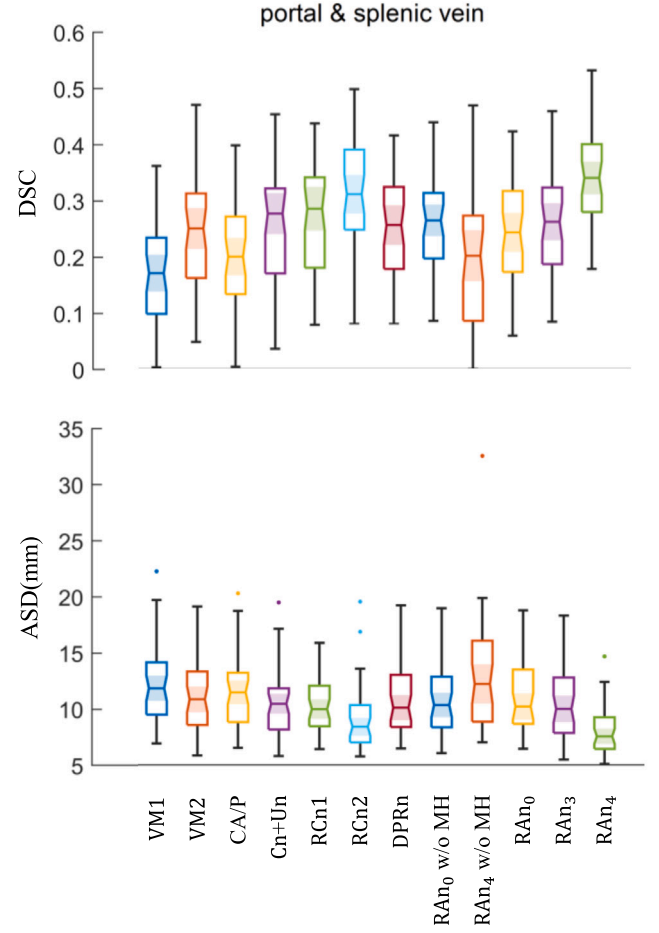
The data used for experiments in this paper are publicly available from <https://zenodo.org/record/3835682> (lung) and https://github.com/ucl-candi/datasets_deepreg_demo/archive/abdct.zip (abd).

Acknowledgments

J.-Q. Z. acknowledges the Kennedy Trust Prize Studentship (AZT00050-AZ04). N.H.L. acknowledges the Centre for OA Pathogenesis Versus Arthritis (Versus Arthritis grant 21621). B.W.P. acknowledges the Rutherford Fund at Health Data Research UK (grant no. MR/S004092/1).

Appendix A. Math

The denotations of symbols are the same as above.

**Fig. B.18.** Results on portal splenic vein.

A.1. Proof of Theorem 1 (Regional Dependency)

The pooling mapping \mathcal{P} from the original full-resolution image's coordinate \mathbf{x} to the k_{th} feature map with pool size p_k is denoted as:

$$\mathcal{P}(\mathbf{x}; p_k) := \lfloor \mathbf{x} / p_k \rfloor, \quad (\text{A.1})$$

which thus satisfy:

$$\exists (\mathbf{x}, \mathbf{y}) \in \{(\mathbf{x}, \mathbf{y}) \mid \|\mathbf{x} - \mathbf{y}\|_\infty < p_k\}, \mathcal{P}(\mathbf{x}; p_k) = \mathcal{P}(\mathbf{y}; p_k) \quad (\text{A.2})$$

As stated in the manuscript, coarse-to-fine registration implies $\forall k \in [1, K) \cap \mathbb{Z}, p_k \geq p_{k+1}, s_k \geq s_{k+1}$, and thus:

$$\forall (\mathbf{x}, \mathbf{y}) \in \{(\mathbf{x}, \mathbf{y}) \mid \|\mathbf{x} - \mathbf{y}\|_\infty \geq p_{k'}\}, \mathcal{P}(\mathbf{x}; p_k) \neq \mathcal{P}(\mathbf{y}; p_k) \quad (\text{A.3})$$

where $k \geq k'$. Because the DDF predicted by k_{th} RA module has the same resolution as feature map:

$$\begin{cases} \phi_k[\mathbf{x}] \equiv \phi_k[\mathbf{y}] & \text{if } \mathcal{P}(\mathbf{x}; p_k) = \mathcal{P}(\mathbf{y}; p_k) \\ \phi_k[\mathbf{x}] \neq \phi_k[\mathbf{y}] & \text{if } \mathcal{P}(\mathbf{x}; p_k) \neq \mathcal{P}(\mathbf{y}; p_k) \end{cases} \quad (\text{A.4})$$

so that:

$$\begin{aligned} &\exists (\mathbf{x}, \mathbf{y}) \in \{(\mathbf{x}, \mathbf{y}) \mid \|\mathbf{x} - \mathbf{y}\|_\infty < p_k\}, \\ &\quad \phi_k[\mathbf{x}] \equiv \phi_k[\mathbf{y}]; \\ &\forall (\mathbf{x}, \mathbf{y}, k) \in \{(\mathbf{x}, \mathbf{y}, k) \mid \|\mathbf{x} - \mathbf{y}\|_\infty \geq p_{k'}, k \geq k'\}, \\ &\quad \phi_k[\mathbf{x}] \neq \phi_k[\mathbf{y}]. \end{aligned} \quad (\text{A.5})$$

Table B.6

Network structures of the decoder for Residual Aligner Networks (RAN₀, RAN₃, RAN₄, RAN₄⁺) with varying channels ($c_0 = 32, 32, 36, 48$; $c_1 = 64, 48, 44, 48$; $c_2 = 48, 48, 44, 48$; $c_3 = 32, 32, 28, 32$; $c_4 = 24, 32, 28, 32$), pooling scales and layer inputs.

Layer(s)	ker	dila	chns	RAN ₀		RAN ₃		RAN ₄		RAN ₄ ⁺		out
				scale	in	scale	in	scale	in	scale	in	
upsample				×		✓		✓		✓		
conv,norm,act	3	1	$c_0/32$	16	s4	2	s4,s3	1	s4,s3,s2	1	s4,s3,s2	r5
conv,norm,act	3	1	32/32	16	r5	2	r5	1	r5	1	r5	f5
conv,norm	3	3	32/32	16	f5	2	f5	1	f5	1	f5	f5
act	–	–	32/32	16	f5+r5	2	f5+r5	1	f5+r5	1	f5+r5	$F_0^{s/t}$
upsample				✓		×		×		×		
conv,norm,act	3	1	$c_1/32$	8	$F_0^{s/t} s4$	2	$F_0^{s/t} s4$	1	$F_0^{s/t} s4$	1	$F_0^{s/t} s4$	r6
conv,norm,act	3	1	32/32	8	r6	2	r6	1	r6	1	r6	f6
conv,norm	3	3	32/32	8	f6	2	f6	1	f6	1	f6	f6
act	–	–	32/32	8	f6+r6	2	f6+r6	1	f6+r6	1	f6+r6	$F_1^{s/t}$
upsample				✓		×		×		×		
conv,norm,act	3	1	$c_2/16$	4	$F_1^{s/t} s3$	2	$F_1^{s/t} s3$	1	$F_1^{s/t} s3$	1	$F_1^{s/t} s3$	r7
conv,norm,act	3	1	16/16	4	r7	2	r7	1	r7	1	r7	f7
conv,norm	3	3	16/16	4	f7	2	f7	1	f7	1	f7	f7
act	–	–	16/16	4	f7+r7	2	f7+r7	1	f7+r7	1	f7+r7	$F_2^{s/t}$
upsample				✓		×		×		×		
conv,norm,act	3	1	$c_3/16$	2	$F_2^{s/t} s2$	2	$F_2^{s/t} s2$	1	$F_2^{s/t} s2$	1	$F_2^{s/t} s2$	r8
conv,norm,act	3	1	16/16	2	r8	2	r8	1	r8	1	r8	f8
conv,norm	3	3	16/16	2	f8	2	f8	1	f8	1	f8	f8
act	–	–	16/16	2	f8+r8	2	f8+r8	1	f8+r8	1	f8+r8	$F_3^{s/t}$
upsample				✓		✓		×		×		
conv,norm,act	3	1	$c_4/8$	1	$F_3^{s/t} s1$	1	$F_3^{s/t} s1$	1	$F_3^{s/t} s1$	1	$F_3^{s/t} s1$	r9
conv,norm,act	3	1	8/8	1	r9	1	r9	1	r9	1	r9	f9
conv,norm	3	3	8/8	1	f9	1	f9	1	f9	1	f9	f9
act	–	–	8/8	1	f9+r9	1	f9+r9	1	f9+r9	1	f9+r9	s9
conv	1	1	8/d	1	s9	1	s9	1	s9	1	s9	$F_4^{s/t}$

According to Eq. (4), $\varphi = \phi_k \circ \phi_{k-1}^{-1}$, the k th predicted DDF ϕ_k can be decomposed as:

$$\begin{aligned}\phi_k[\mathbf{x}] &= \varphi_k \circ \phi_{k-1}[\mathbf{x}] \\ &= \varphi_k + \phi_{k-1}[\mathbf{x} - \varphi_k[\mathbf{x}]] \\ &= \phi_k \circ \phi_{k-1}^{-1}[\mathbf{x}] + \phi_{k-1}[\mathbf{x} - \phi_k \circ \phi_{k-1}^{-1}[\mathbf{x}]]\end{aligned}\quad (\text{A.6})$$

where $\phi_k \circ \phi_{k-1}^{-1}$ is regressed by \mathcal{R}_k and \mathcal{T}_k in RA module as described in the manuscript. The difference between two displacements $\phi_k[\mathbf{x}]$ and $\phi_k[\mathbf{y}]$ can be written as:

$$\begin{aligned}\Delta\phi_k(\mathbf{x}, \mathbf{y}) &:= \phi_k[\mathbf{x}] - \phi_k[\mathbf{y}] \\ &= \phi_k \circ \phi_{k-1}^{-1} \circ \phi_{k-1}[\mathbf{x}] - \phi_k \circ \phi_{k-1}^{-1} \circ \phi_{k-1}[\mathbf{y}] \\ &= (\phi_k \circ \phi_{k-1}^{-1}[\mathbf{x}] + \phi_{k-1}[\mathbf{x} - \phi_k \circ \phi_{k-1}^{-1}[\mathbf{x}]] \\ &\quad - (\phi_k \circ \phi_{k-1}^{-1}[\mathbf{y}] + \phi_{k-1}[\mathbf{y} - \phi_k \circ \phi_{k-1}^{-1}[\mathbf{y}]]) \\ &= \underbrace{\phi_k \circ \phi_{k-1}^{-1}[\mathbf{x}] - \phi_k \circ \phi_{k-1}^{-1}[\mathbf{y}]}_{\text{i}} \\ &\quad + \underbrace{\phi_{k-1}[\mathbf{x} - \phi_k \circ \phi_{k-1}^{-1}[\mathbf{x}]] - \phi_{k-1}[\mathbf{y} - \phi_k \circ \phi_{k-1}^{-1}[\mathbf{y}]]}_{\text{ii}}\end{aligned}\quad (\text{A.7})$$

where the range of Eq. (A.7)(i) is limited by the k th RA module and Eq. (A.7)(ii) can be substituted with $\phi_{k-1}[\mathbf{x}'] - \phi_{k-1}[\mathbf{y}']$ by $\mathbf{x}' := \mathbf{x} - \phi_k \circ \phi_{k-1}^{-1}[\mathbf{x}]$, $\mathbf{y}' := \mathbf{y} - \phi_k \circ \phi_{k-1}^{-1}[\mathbf{y}]$, where the iterative equation of Eq. (A.7) can be thus written as:

$$\begin{cases} \Delta\phi_k(\mathbf{x}^k, \mathbf{y}^k) - (\mathbf{x}^k - \mathbf{y}^k) = \Delta\phi_{k-1}(\mathbf{x}^{k-1}, \mathbf{y}^{k-1}) - (\mathbf{x}^{k-1} - \mathbf{y}^{k-1}) \\ \mathbf{x}^k - \phi_k \circ \phi_{k-1}^{-1}[\mathbf{x}^k] = \mathbf{x}^{k-1} \\ \mathbf{y}^k - \phi_k \circ \phi_{k-1}^{-1}[\mathbf{y}^k] = \mathbf{y}^{k-1} \end{cases}\quad (\text{A.8})$$

starting from $\Delta\phi_0(\mathbf{x}^0, \mathbf{y}^0) := 0 \forall \mathbf{x}^0, \mathbf{y}^0 \in \mathbb{Z}^d$. The analytic equation is derived as:

$$\begin{cases} \Delta\phi_k(\mathbf{x}^k, \mathbf{y}^k) = \Delta\phi_{k'-1}(\mathbf{x}^{k'-1}, \mathbf{y}^{k'-1}) + \sum_{k'=k''+1}^k \Delta\varphi_{k'}(\mathbf{x}^{k'}, \mathbf{y}^{k'}) \\ \mathbf{x}^k - \mathbf{y}^k = (\mathbf{x}^{k'-1} - \mathbf{y}^{k'-1}) + \sum_{k'=k''+1}^k \Delta\varphi_{k'}(\mathbf{x}^{k'}, \mathbf{y}^{k'}) \end{cases}\quad (\text{A.9})$$

where $\Delta\varphi_k$ is defined as:

$$\begin{aligned}\Delta\varphi_k(\mathbf{x}, \mathbf{y}) &:= \varphi_k[\mathbf{x}] - \varphi_k[\mathbf{y}] \\ &= \phi_k \circ \phi_{k-1}^{-1}[\mathbf{x}] - \phi_k \circ \phi_{k-1}^{-1}[\mathbf{y}]\end{aligned}\quad (\text{A.10})$$

Substitute Eq. (A.5) into Eq. (A.9), we can conclude that

$$\begin{aligned}\exists (\mathbf{x}^k, \mathbf{y}^k) \in & \\ \{(\mathbf{x}, \mathbf{y}) | \|\mathbf{x} - \mathbf{y}\|_\infty < p_{k''-1} + \sum_{k'=k''+1}^k \Delta\varphi_{k'}(\mathbf{x}^{k'}, \mathbf{y}^{k'})\}, & \\ \Delta\phi_k(\mathbf{x}^k, \mathbf{y}^k) = \sum_{k'=k''+1}^k \Delta\varphi_{k'}(\mathbf{x}^{k'}, \mathbf{y}^{k'}); & \\ \forall (\mathbf{x}^k, \mathbf{y}^k, k) \in & \\ \{(\mathbf{x}, \mathbf{y}, k) | \|\mathbf{x} - \mathbf{y}\|_\infty \geq p_{k''} + \sum_{k'=k''+1}^k \Delta\varphi_{k'}(\mathbf{x}^{k'}, \mathbf{y}^{k'}), k \geq k''\}, & \\ \Delta\phi_k(\mathbf{x}^k, \mathbf{y}^k) \geq \sum_{k'=k''+1}^k \Delta\varphi_{k'}(\mathbf{x}^{k'}, \mathbf{y}^{k'}); & \end{aligned}$$

with satisfying $\sup(\Delta\phi_k(\mathbf{x}^k, \mathbf{y}^k)) = \sup(\|\phi_k[\mathbf{x}] - \phi_k[\mathbf{y}]\|_\infty)$, $\sup(\Delta\varphi_k(\mathbf{x}^k, \mathbf{y}^k)) = 2a_k$, which thus prove **Theorem 1 (Regional Dependency)**.

A.2. Confidence-weighted interpolation

The areas lacking texture or structural features usually result in a deviation in prediction and thus require correction from the interpolation or smoothing based on the neighboring predicted values. To strengthen the different displacement at each pixel/voxel with individual weights, the confidence values are respectively quantified by C^1 for $\hat{\phi}_k$ and ϕ_{k-1} .

For an example of a simple Gaussian-based smoothing on the $(k-1)$ th-level DDF ϕ_{k-1} adaptively weighted by a confidence map C :

$$\begin{aligned}\text{smooth}(\phi_{k-1}, C) &= \overbrace{C \odot (\phi_{k-1} * \mathbf{G})}^{\text{smoothed DDF}} + \overbrace{(1-C) \odot \phi_{k-1}}^{\text{original DDF}} \\ &= \phi_{k-1} - \underbrace{C \odot (\phi_{k-1} * \mathbf{L})}_{=\phi_{k-1}'}\end{aligned}\quad (\text{A.11})$$

where \mathbf{G} denotes a Gaussian filter kernel for smoothing, $\mathbf{L} := 1 - \mathbf{G}$ denotes the Laplacian filter kernel. Here the Laplacian convolution $(\phi_{k-1} * \mathbf{L})$ is regressed by $C^2(\phi_{k-1})$, and the confidence weight $C := C^1(\theta_k)$.

Table B.7Network structure of Residual Aligner (RA) modules for RANs (RAN₀, RAN₃, RAN₄, RAN₄⁺) with varying pooling scales and dilation rates.

Layer(s)	ker	chns	RAN ₀		RAN ₃		RAN ₄		RAN ₄ ⁺		in	out
			scale	dila	scale	dila	scale	dila	scale	dila		
conv,act,conv,act	3	64/18/18	16	1	2	8	1	16	1	16	$F_0^s F_0^t$	m0
conv,act,conv	3	18/27/md	16	1	2	1	1	1	1	1	m0	$\hat{\phi}_0$
conv,act,conv	3	18/18/m	16	1	2	1	1	1	1	1	m0	θ_0
conv,norm,act	3	18/1	16	1	2	1	1	1	1	1	θ_0	θ_0'
conv	1	m/m	16	0	2	0	1	0	1	0	θ_0	θ_0
reshape,conv	3	md/m/9	16	1	2	1	1	1	1	1	$\sigma(\theta_0) \odot \hat{\phi}_0$	df0
conv,reshape	3	9/1/d	16	1	2	1	1	1	1	1	$\theta_0' \odot df0$	ϕ_0
upsample			✓		×		×		×		ϕ_0, θ_0	ϕ_0, θ_0
conv,act,conv,act	3	64/18/18	8	1	2	4	1	8	1	8	$\phi_0(F_1^s) F_1^t$	m1
conv,act,conv	3	18/27/md	8	1	2	1	1	1	1	1	m1	$\hat{\phi}_1$
conv,act,conv	3	18/18/m	8	1	2	1	1	1	1	1	m1	θ_1
conv,norm,act	3	18/1	8	1	2	1	1	1	1	1	θ_1, θ_0	θ_1', θ_0'
conv	1	2m/m	8	0	2	0	1	0	1	0	$\theta_1 \theta_0$	θ_1
reshape,conv	3	md/m/9	8	1	2	1	1	1	1	1	$\sigma(\theta_1) \odot \hat{\phi}_1$	df1
reshape,conv	3	md/m/9	8	1	2	1	1	1	1	1	$\sigma(\theta_1) \otimes \phi_0$	dp1
conv,reshape	3	18/1/d	8	1	2	1	1	1	1	1	$\theta_1' \odot df1 \theta_0' \odot dp1$	ϕ_1
upsample			✓		×		×		×		ϕ_1, θ_1	ϕ_1, θ_1
conv,act,conv,act	3	32/18/18	4	1	2	2	1	4	1	4	$\phi_1(F_2^s) F_2^t$	m2
conv,act,conv	3	18/27/md	4	1	2	1	1	1	1	1	m2	$\hat{\phi}_2$
conv,act,conv	3	18/18/m	4	1	2	1	1	1	1	1	m2	θ_2
conv,norm,act	3	18/1	4	1	2	1	1	1	1	1	θ_2, θ_1	θ_2', θ_1'
conv	1	2m/m	4	0	2	0	1	0	1	0	$\theta_2 \theta_1$	θ_2
reshape,conv	3	md/m/9	4	1	2	1	1	1	1	1	$\sigma(\theta_2) \odot \hat{\phi}_2$	df2
reshape,conv	3	md/m/9	4	1	2	1	1	1	1	1	$\sigma(\theta_2) \otimes \phi_1$	dp2
conv,reshape	3	18/1/d	4	1	2	1	1	1	1	1	$\theta_2' \odot df2 \theta_1' \odot dp2$	ϕ_2
upsample			✓		×		×		×		ϕ_2, θ_2	ϕ_2, θ_2
conv,act,conv,act	3	32/18/18	2	1	2	1	1	2	1	2	$\phi_2(F_3^s) F_3^t$	m3
conv,act,conv	3	18/27/md	2	1	2	1	1	1	1	1	m3	$\hat{\phi}_3$
conv,act,conv	3	18/18/m	2	1	2	1	1	1	1	1	m3	θ_3
conv,norm,act	3	18/1	2	1	2	1	1	1	1	1	θ_3, θ_2	θ_3', θ_2'
conv	1	2m/m	2	0	2	0	1	0	1	0	$\theta_3 \theta_2$	θ_3
reshape,conv	3	md/m/9	2	1	2	1	1	1	1	1	$\sigma(\theta_3) \odot \hat{\phi}_3$	df3
reshape,conv	3	md/m/9	2	1	2	1	1	1	1	1	$\sigma(\theta_3) \otimes \phi_2$	dp3
conv,reshape	3	18/1/d	2	1	2	1	1	1	1	1	$\theta_3' \odot df3 \theta_2' \odot dp3$	ϕ_3
upsample			✓		✓		×		×		ϕ_3, θ_3	ϕ_3, θ_3
conv,act,conv,act	3	16/18/18	1	1	1	1	1	1	1	1	$\phi_3(F_4^s) F_4^t$	m4
conv,act,conv	3	18/27/md	1	1	1	1	1	1	1	1	m4	$\hat{\phi}_4$
conv,act,conv	3	18/18/m	1	1	1	1	1	1	1	1	m4	θ_4
conv,norm,act	3	18/1	1	1	1	1	1	1	1	1	θ_4, θ_3	θ_4', θ_3'
conv	1	2m/m	1	0	1	0	1	0	1	0	$\theta_4 \theta_3$	θ_4
reshape,conv	3	md/m/9	1	1	1	1	1	1	1	1	$\sigma(\theta_4) \odot \hat{\phi}_4$	df4
reshape,conv	3	md/m/9	1	1	1	1	1	1	1	1	$\sigma(\theta_4) \otimes \phi_3$	dp4
conv,reshape	3	18/1/d	1	1	1	1	1	1	1	1	$\theta_4' \odot df4 \theta_3' \odot dp4$	ϕ

is implicitly regressed from θ_k with general representation for the aim of higher accuracy. Thus the calculation of $\text{smooth}(\phi_{k-1}, C^1(\theta_{k-1}))$ and $\text{smooth}(\hat{\phi}_k, C^1(\theta_k))$ could be regressed by $C^4([\varphi'_k, \hat{\phi}_k, \phi'_{k-1}, \phi_{k-1}])$ in Eq. (12),

A.3. Multi-head disentanglement

To disentangle the predicted DDF with preserving discontinuities and the trend of motions, the M-H masks $\mathbf{M} := \text{softmax}(\theta_k)$ is inserted into Eq. (A.11) for decoupling and smoothing the prediction on the different regions of DDF ϕ_k :

$$\text{smooth}(\phi_{k-1}, \mathbf{C}, \mathbf{M}) = \phi_{k-1} - \sum_{\{m\}} \underbrace{\mathbf{C} \odot ((\mathbf{M} \otimes \phi_{k-1}) * \mathbf{L})}_{=\phi'_{k-1}} \quad (\text{A.12})$$

and M-H residual DDF $\hat{\phi}_k$:

$$\text{smooth}(\hat{\phi}_k, \mathbf{C}, \mathbf{M}) = \sum_{\{m\}} \phi_k - \underbrace{\mathbf{C} \odot ((\mathbf{M} \otimes \hat{\phi}_k) * \mathbf{L})}_{=\varphi'_k} \quad (\text{A.13})$$

where $\sum_{\{m\}}$ denotes the head-dimension sum. The calculation of Eqs. (A.12) and (A.13) could be regressed by:

$$\phi_k = C^4([\varphi'_k, \sum_{\{m\}} (\hat{\phi}_k), \phi'_{k-1}, \phi_{k-1}]) \quad (\text{A.14})$$

in Eq. (12) to predict the output DDF of the k th RA module ϕ_k .

Appendix B. Network architecture

The MS network structure details of the encoder, the decoder, and the RA modules are respectively illustrated in Tables B.5–B.7, including RAN₀, RAN₃, RAN₄ and RAN₄⁺.

Appendix C. Additional results

We compared RAN with the relevant state-of-the-art network structures. The Voxelmorph (Balakrishnan et al., 2019) (VM1/VM2: light-/heavy-weight model) is adopted as the representative method of DR. The composite network combining CNN (Cn: Global-net) and U-net (Un: Local-net) following to Hu et al. (2018), as well as 5-RCn (Zhao et al., 2019a) (RCn1/RCn2: light-/heavy-weight model) are also adopted into the framework as the relevant baselines representing multi-stage Cascaded (Cas) networks. DPRn (Kang et al., 2022) is selected as the baseline for FP networks. Additionally, we also replace RA module with cross attention (Attn) (Vaswani et al., 2017) to compare the performance at the module level.

To clearly show the performance detail of the previous relevant models compared with our RANs as well as the ablation studies on

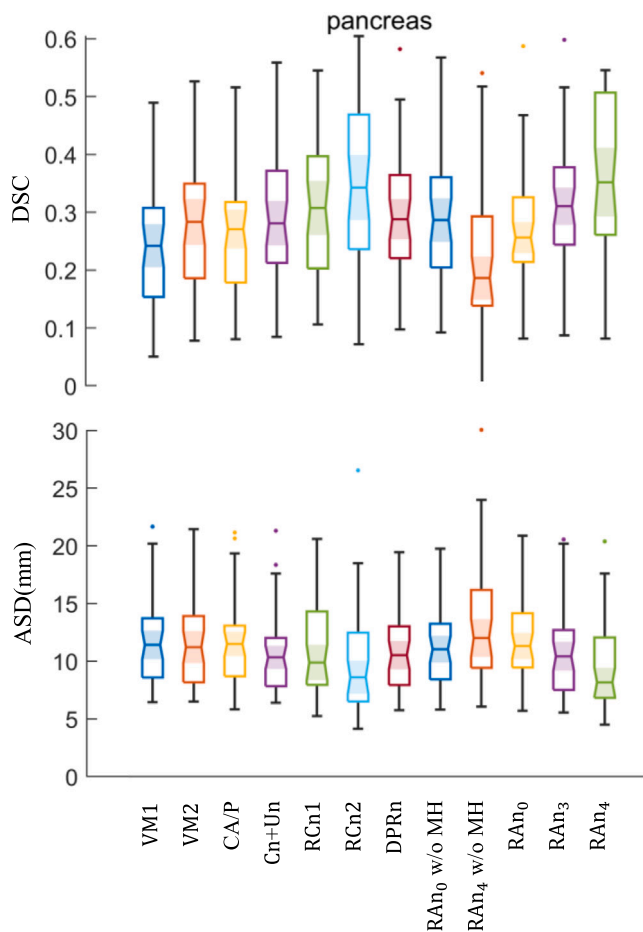


Fig. B.19. Results on pancreas.

the nine organs: spleen (Fig. B.11), right kidney (Fig. B.12), left kidney (Fig. B.13), esophagus (Fig. B.14), liver (Fig. B.15), aorta (Fig. B.16), inferior vena cava (Fig. B.17), portal splenic vein (Fig. B.18), and pancreas (Fig. B.19).

References

- Balakrishnan, G., Zhao, A., Sabuncu, M.R., Guttag, J., Dalca, A.V., 2018. An unsupervised learning model for deformable medical image registration. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 9252–9260.
- Balakrishnan, G., Zhao, A., Sabuncu, M.R., Guttag, J., Dalca, A.V., 2019. VoxelMorph: a learning framework for deformable medical image registration. *IEEE Trans. Med. Imaging* 38 (8), 1788–1800.
- Cao, Y., Zhu, Z., Rao, Y., Qin, C., Lin, D., Dou, Q., Ni, D., Wang, Y., 2021. Edge-aware pyramidal deformable network for unsupervised registration of brain MR images. *Front. Neurosci.* 14, 1464.
- Chang, C.-H., Chou, C.-N., Chang, E.Y., 2017. Clkn: Cascaded lucas-kanade networks for image alignment. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2213–2221.
- Chen, J., Frey, E.C., He, Y., Segars, W.P., Li, Y., Du, Y., 2022. Transmorph: Transformer for unsupervised medical image registration. *Med. Image Anal.* 82, 102615.
- Chen, J., He, Y., Frey, E.C., Li, Y., Du, Y., 2021a. Vit-v-net: Vision transformer for unsupervised volumetric medical image registration. *arXiv preprint arXiv:2104.06468*.
- Chen, X., Xia, Y., Ravikumar, N., Frangi, A.F., 2021b. A deep discontinuity-preserving image registration network. In: Medical Image Computing and Computer Assisted Intervention—MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part IV 24. Springer, pp. 46–55.
- Fischer, B., Modersitzki, J., 2008. Ill-posed medicine—an introduction to image registration. *Inverse Probl.* 24 (3), 034008.
- Fu, Y., Brown, N.M., Saeed, S.U., Casamitjana, A., Delaunay, R., Yang, Q., Greenwood, A., Min, Z., Blumberg, S.B., Iglesias, J.E., et al., 2020. DeepReg: a deep learning toolkit for medical image registration. *J. Open Source Softw.* 5 (55), 2705.

- He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778.
- Heinrich, M.P., 2019. Closing the gap between deep and conventional image registration using probabilistic dense displacement networks. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 50–58.
- Heinrich, M.P., Jenkinson, M., Papież, B.W., Glesson, F.V., Brady, M., Schnabel, J.A., 2013. Edge-and detail-preserving sparse image representations for deformable registration of chest MRI and CT volumes. In: International Conference on Information Processing in Medical Imaging. Springer, pp. 463–474.
- Heise, P., Klose, S., Jensen, B., Knoll, A., 2013. Pm-huber: Patchmatch with huber regularization for stereo matching. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2360–2367.
- Hering, A., Murphy, K., van Ginneken, B., 2020. Learn2Reg challenge: CT lung registration - training data. <http://dx.doi.org/10.5281/zenodo.3835682>.
- Hering, A., et al., 2021. Learn2Reg: comprehensive multi-task medical image registration challenge, dataset and evaluation in the era of deep learning. *arXiv:2112.04489*.
- Hu, Y., Modat, M., Gibson, E., Li, W., Ghavami, N., Bonmati, E., Wang, G., Bandula, S., Moore, C.M., Emberton, M., et al., 2018. Weakly-supervised convolutional neural networks for multimodal image registration. *Med. Image Anal.* 49, 1–13.
- Hua, R., Pozo, J.M., Taylor, Z.A., Frangi, A.F., 2017. Multiresolution eXtended Free-Form Deformations (XFFD) for non-rigid registration with discontinuous transforms. *Med. Image Anal.* 36, 113–122.
- Jaderberg, M., Simonyan, K., Zisserman, A., et al., 2015. Spatial transformer networks. *Adv. Neural Inf. Process. Syst.* 28, 2017–2025.
- Kang, M., Hu, X., Huang, W., Scott, M.R., Reyes, M., 2022. Dual-stream pyramid registration network. *Med. Image Anal.* 78, 102379.
- Li, Z., Liu, X., Drenkow, N., Ding, A., Creighton, F.X., Taylor, R.H., Unberath, M., 2021. Revisiting stereo depth estimation from a sequence-to-sequence perspective with transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6197–6206.
- Liu, R., Li, Z., Fan, X., Zhao, C., Huang, H., Luo, Z., 2021. Learning deformable image registration from optimization: perspective, modules, bilevel training and beyond. *IEEE Trans. Pattern Anal. Mach. Intell.*
- Lv, Z., Dellaert, F., Reh, J.M., Geiger, A., 2019. Taking a deeper look at the inverse compositional algorithm. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4581–4590.
- Mok, T.C., Chung, A., 2020. Fast symmetric diffeomorphic image registration with convolutional neural networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4644–4653.
- Ng, E., Ebrahimi, M., 2020. An unsupervised learning approach to discontinuity-preserving image registration. In: Biomedical Image Registration: 9th International Workshop, WBIR 2020, Portorož, Slovenia, December 1–2, 2020, Proceedings 9. Springer, pp. 153–162.
- Papież, B.W., Franklin, J.M., Heinrich, M.P., Gleeson, F.V., Brady, M., Schnabel, J.A., 2018. GIFTed Demons: deformable image registration with local structure-preserving regularization using supervoxels for liver applications. *J. Med. Imaging* 5 (2), 024001.
- Papież, B.W., Heinrich, M.P., Fehrenbach, J., Risser, L., Schnabel, J.A., 2014. An implicit sliding-motion preserving regularisation via bilateral filtering for deformable image registration. *Med. Image Anal.* 18 (8), 1299–1311.
- Ranjan, A., Black, M.J., 2017. Optical flow estimation using a spatial pyramid network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4161–4170.
- Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 234–241.
- Rueckert, D., Sonoda, L.I., Hayes, C., Hill, D.L., Leach, M.O., Hawkes, D.J., 1999. Nonrigid registration using free-form deformations: application to breast MR images. *IEEE Trans. Med. Imaging* 18 (8), 712–721.
- Schmidt-Richberg, A., Werner, R., Handels, H., Ehrhardt, J., 2012. Estimation of slipping organ motion by registration with direction-dependent regularization. *Med. Image Anal.* 16 (1), 150–159.
- Shen, Z., Han, X., Xu, Z., Niethammer, M., 2019. Networks for joint affine and non-parametric image registration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4224–4233.
- Song, X., Guo, H., Xu, X., Chao, H., Xu, S., Turkbey, B., Wood, B.J., Wang, G., Yan, P., 2021. Cross-modal attention for MRI and ultrasound volume registration. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 66–75.
- Sotiras, A., Davatzikos, C., Paragios, N., 2013. Deformable medical image registration: A survey. *IEEE Trans. Med. Imaging* 32 (7), 1153–1190.
- Sun, J., Shen, Z., Wang, Y., Bao, H., Zhou, X., 2021. LoFTR: Detector-free local feature matching with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8922–8931.
- Sun, D., Yang, X., Liu, M.-Y., Kautz, J., 2018. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 8934–8943.

- Thirion, J.-P., 1998. Image matching as a diffusion process: an analogy with Maxwell's demons. *Med. Image Anal.* 2 (3), 243–260.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. In: *Advances in Neural Information Processing Systems*. pp. 5998–6008.
- Vercauteren, T., Pennec, X., Perchant, A., Ayache, N., 2008. Diffeomorphic demons using ITK's finite difference solver hierarchy. <http://dx.doi.org/10.54294/ux2obj>.
- Vercauteren, T., Pennec, X., Perchant, A., Ayache, N., 2009. Diffeomorphic demons: Efficient non-parametric image registration. *NeuroImage* 45 (1), S61–S72.
- Vishnevskiy, V., Gass, T., Szekely, G., Tanner, C., Goksel, O., 2016. Isotropic total variation regularization of displacements in parametric image registration. *IEEE Trans. Med. Imaging* 36 (2), 385–395.
- de Vos, B.D., Berendsen, F.F., Viergever, M.A., Sokooti, H., Staring, M., Išgum, I., 2019. A deep learning framework for unsupervised affine and deformable image registration. *Med. Image Anal.* 52, 128–143.
- Wendland, H., 1995. Piecewise polynomial, positive definite and compactly supported radial functions of minimal degree. *Adv. Comput. Math.* 4, 389–396.
- Xiao, J., Cheng, H., Sawhney, H., Rao, C., Isnardi, M., 2006. Bilateral filtering-based optical flow estimation with occlusion detection. In: *European Conference on Computer Vision*. Springer, pp. 211–224.
- Xu, Z., Luo, J., Yan, J., Li, X., Jayender, J., 2021. F3RNet: full-resolution residual registration network for deformable image registration. *Int. J. Comput. Assist. Radiol. Surg.* 16 (6), 923–932.
- Zhang, Y., Pei, Y., Zha, H., 2021. Learning dual transformer network for diffeomorphic registration. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, pp. 129–138.
- Zhao, S., Dong, Y., Chang, E.I., Xu, Y., et al., 2019a. Recursive cascaded networks for unsupervised medical image registration. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 10600–10610.
- Zhao, S., Lau, T., Luo, J., Eric, I., Chang, C., Xu, Y., 2019b. Unsupervised 3D end-to-end medical image registration with volume tweening network. *IEEE J. Biomed. Health Inf.* 24 (5), 1394–1404.
- Zheng, J.-Q., Lim, N.H., Papież, B.W., 2023. Accurate volume alignment of arbitrarily oriented tibiae based on a mutual attention network for osteoarthritis analysis. *Comput. Med. Imaging Graph.* 106, 102204.
- Zheng, J.-Q., Wang, Z., Huang, B., Vincent, T., Lim, N.H., Papież, B.W., 2022. Recursive deformable image registration network with mutual attention. In: *Annual Conference on Medical Image Understanding and Analysis*. Springer, pp. 75–86.
- Zhou, X.-Y., Zheng, J.-Q., Li, P., Yang, G.-Z., 2020. ACNN: a full resolution dcnn for medical image segmentation. In: *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, pp. 8455–8461.