Modality-Agnostic Structural Image Representation Learning for Deformable Multi-Modality Medical Image Registration

Tony C. W. Mok[1,2*]   Zi Li[1,2*]   Yunhao Bai[1]   Jianpeng Zhang[1,2,4]   Wei Liu[1,2]

Yan-Jie Zhou[1,2,4]   Ke Yan[1,2]   Dakai Jin[1]   Yu Shi[3]   Xiaoli Yin[3]   Le Lu[1]   Ling Zhang[1]

[1] DAMO Academy, Alibaba Group

[2] Hupan Lab, 310023, Hangzhou, China

[3] Shengjing Hospital of China Medical University, China

[4] College of Computer Science and Technology, Zhejiang University, China

cwmokab@connect.ust.hk

Abstract

Establishing dense anatomical correspondence across distinct imaging modalities is a foundational yet challenging procedure for numerous medical image analysis studies and image-guided radiotherapy. Existing multi-modality image registration algorithms rely on statistical-based similarity measures or local structural image representations. However, the former is sensitive to locally varying noise, while the latter is not discriminative enough to cope with complex anatomical structures in multimodal scans, causing ambiguity in determining the anatomical correspondence across scans with different modalities. In this paper, we propose a modality-agnostic structural representation learning method, which leverages Deep Neighbourhood Self-similarity (DNS) and anatomy-aware contrastive learning to learn discriminative and contrast-invariance deep structural image representations (DSIR) without the need for anatomical delineations or pre-aligned training images. We evaluate our method on multiphase CT, abdomen MR-CT, and brain MR T1w-T2w registration. Comprehensive results demonstrate that our method is superior to the conventional local structural representation and statistical-based similarity measures in terms of discriminability and accuracy.

1. Introduction

Determining anatomical correspondence between multimodal data is crucial for medical image analysis and clinical applications, including diagnostic settings [33], surgical planning [1, 57] and post-operative evaluation [39]. As a vital component for modern medical image analysis
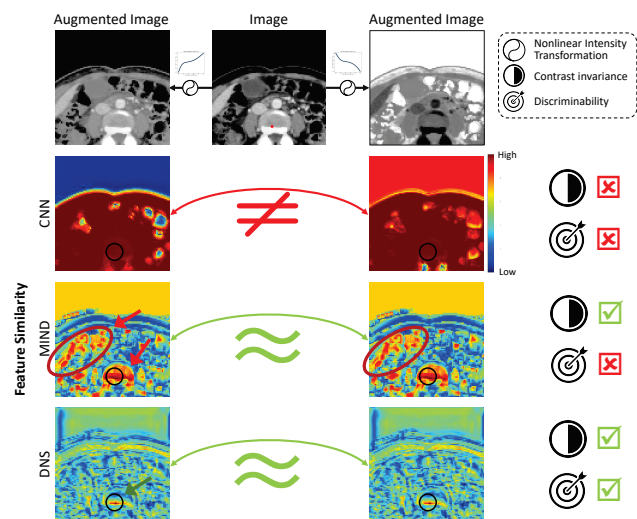

Figure 1. Visualization of feature similarity between the marked feature vector (red dot) of the image and all feature vectors of augmented images using the convolutional neural network without pertaining (CNN), Modality Independent Neighbourhood Descriptor (MIND), and our proposed Deep Neighbourhood Self-similarity (DNS). Our method captures the contrast invariant and high discriminability structural representation of the image, reducing the ambiguity in matching the anatomical correspondence between multimodal images.

studies and image-guided interventions, deformable multimodal registration aims to establish the dense anatomical correspondence between multimodal scans and fuse information from multimodal scans, e.g., propagating anatomical or tumour delineation for image-guided radiotherapy [30]. Since different imaging modalities provide valuable complementary visual cues and diagnosis information of the patient, precise anatomical alignment between multimodal scans benefits the radiological observation and the

---
*Contributed equally.

subsequent downstream computerized analyses. However, finding anatomical correspondences between homologous points in multimodal images is notoriously challenging due to the complex appearance changes across modalities. For instance, in multiphase abdomen computed tomography (CT) scans, the soft tissues can be deformed due to gravity, body motion, and other muscle contractions, resulting in an unavoidable large non-linear misalignment between subsequent imaging scans. Moreover, anatomical structures and tumours in multiphase CT scans show heterogeneous intensity distribution across different multiphase contrast-enhanced CT scans due to the intravenously injected contrast agent during the multiphase CT imaging.

Despite there being vast research studies [7, 11, 23, 28, 29, 35, 39, 50] on deformable image registration, most of these are focused on the mono-modal registration settings and rely on intensity-based similarity metrics, *i.e.*, normalized cross-correlation (NCC) and mean squared error (MSE), which are not applicable to the multimodal registration. Recently, several methods have proposed to learn an inter-domain similarity metric using supervised learning with pre-aligned training images [14, 25, 44]. However, the perfectly aligned images and the ideal ground truth deformations are often absent in multimodal medical images, which limits the applicability of these methods.

Historically, a pioneering work of Maes *et al.* [32] uses mutual information (MI) [55] to perform rigid multimodal registration. Nevertheless, for deformable multimodal registration, many disadvantages have been identified when using the MI-based similarity measures [45]. Specifically, MI-based similarity measures are sensitive to locally varying noise distribution but not sensitive to the subtle anatomical and vascular structures due to the statistical nature of MI.

As an alternative to directly assessing similarity or MI on the original images, structural image representation approaches have gained great interest for deformable multimodal registration. By computing the intermediate structural image representation independent of the underlying image acquisition, well-established monomodal optimization techniques can be employed to address the multimodal registration problem. A prominent example is the Modality-Independent Neighbourhood Descriptor [17], which is motivated by image self-similarity [48] and able to capture the internal geometric layouts of local self-similarities within images. Yet, such local feature descriptors are not expressive and discriminative enough to cope with complex anatomical structures in abdomen CT, exhibiting many local optima, as shown in Fig. 1. Therefore, it is often jointly used with a dedicated optimization strategy or requires robustness initialization.

In this paper, we analyze and expose the limitations of self-similarity-based feature descriptors and mutual information-based methods in multi-modality registration.

We depart from the classical self-similarity descriptor and propose a novel structural image representation learning paradigm dedicated to learning expressive deep structural image representations (DSIRs) for deformable multimodal registration. Our proposed method reduces the multimodal registration problem to a monomodal one, in which existing well-established monomodal registration algorithms can be applied. To the best of our knowledge, this is the first modality-agnostic structural representation learning approach that learns to capture DSIR with high discriminability from multimodal images without using perfectly aligned image pairs or anatomical delineation.

The main contributions of this work are as follows:

- we propose a novel self-supervised deep structural representation learning approach for multimodal image registration that learns to extract deep structural image representation from standalone medical images, circumventing the need for anatomical delineations or perfectly aligned training image pair for supervision.
- we introduce the Deep Neighbour Self-similarity (DNS), which can capture long-range and complex structural information from medical images, addressing the ambiguity in classical feature descriptors and similarity metrics.
- we propose a novel contrastive learning strategy with non-linear intensity transformation, maximizing the discriminability of the feature representation across anatomical positions with homogeneous and heterogeneous intensity distribution.
- we demonstrate that the proposed deep structural image representation can be adapted to a variety of well-established learning-based and iterative optimization registration algorithms, reducing the multimodal registration problem to a monomodal registration problem.

We rigorously evaluate the proposed method on three challenging multimodal registration tasks: liver multiphase CT registration, abdomen magnetic resonance imaging (MR) to CT registration, and brain MR T1w-T2w registration. Results demonstrate that our method is capable of computing highly expressive and discriminative structural representations of multimodal images, reaching the state-of-the-art performance of conventional methods solely with a simple gradient decent-based optimization framework.

2. Related Work

Multi-modal image registration. In general, multimodal registration methods can be divided into three categories: statistical-based, structural representation-based and deep learning-based methods.

Prominent examples of statistical-based methods use information theory and optimize joint voxel statistics, such as minimizing the MI or normalized mutual information (NMI) as similarity measures for multi-modal registration [3, 4, 52]. These approaches aim to estimate a solution

that minimizes the entropy of the joint histogram between image pairs for rigid registration. However, as evidenced in [45], MI-based similarity measures are restricted to measuring the statistical co-occurrence of image intensities, which are not extendable to the co-occurrence of complex patterns, such as subtle structural information of soft tissues and vascular structures in medical images, and are sensitive to locally varying noise distribution, which is not ideal for deformable multi-modal registration.

To circumvent the limitation of statistical-based methods, another common strategy is to project multimodal images into common intermediate structural representations or measure the misalignment through the modality-invariant similarity metrics [12, 27, 29]. Prominent examples include MIND [17, 18, 20], attribute vectors [49] and Linear Correlation of Linear Combination (LC^2) [56] that are designed to capture a dense structural image representation of multimodal images that are independent of the underlying image acquisition. However, these local structural representations are often not sufficiently discriminative and expressive to drive a non-rigid registration with many degrees of freedom, not differentiable or expensive to compute. As such, these local structural representations are often used in conjunction with robust optimization methods, *e.g.* discrete optimization [19] and Bound Optimization by Quadratic Approximation [46], or requiring robust rigid initialization.

Deep learning-based image registration (DLIR) methods have demonstrated remarkable results on diverse monomodal and multi-modal registration tasks, as evidenced by tremendous registration benchmarks [10, 22]. However, the success of recent DLIR approaches has largely been fueled by the supervision of anatomical segmentation labels [42, 60] or the supervision of perfectly aligned multimodal images [14, 25, 44]. The absence of pre-aligned multimodal medical images and dependence on segmentation labels further restricts their generalizability across different anatomies or modalities. In contrast to the mainstream DLIR and learning-based structural representation methods, our proposed method is fully self-supervised, which circumvents the need for anatomical delineations or perfectly aligned multi-modal images.

Contrastive learning in image registration. Motivated by the success of contrastive learning in visual representation learning [8, 15, 21], several methods [5, 26, 58] adopt contrastive learning to extract anatomical structural embedding for monomodal registration. Contrastive learning focuses on extracting discriminative representations by contrasting positive and negative pairs of instances [16]. These methods use noise contrastive estimation (NCE) [54], learning an anatomical structural representation where the feature vectors from the same anatomical location are brought together, in contrast to feature vectors from differ-

ent anatomical locations. Nevertheless, the learned anatomical structural representations of these methods are not contrast invariance, hence, incapable of multimodal registration. Apart from learning structural representation, a recent work [9] directly minimizes the PatchNCE [43] loss for brain MR T1w-T2w registration. Yet, minimizing the PatchNCE is identical to maximizing a lower bound on mutual information between corresponding spatial locations in the feature maps, which inherits the limitation of MI in deformable registration.

3. Method

3.1. Problem Setup and Overview

Let $\mathbf{F}$, $\mathbf{M}$ be fixed and moving volumes defined over a n-D mutual spatial domain $\Omega \subseteq \mathbb{R}^n$. For simplicity, we further assume that $\mathbf{F}$ and $\mathbf{M}$ are three-dimensional, single-channel, and grayscale images, *i.e.*, $n = 3$ and $\Omega \subseteq \mathbb{R}^3$. In this paper, we aim to extract DSIRs of $\mathbf{F}$ and $\mathbf{M}$, *i.e.*, $\mathbf{D_F}$ and $\mathbf{D_M}$, with high discriminability such that only the cosine similarity of two feature vectors in the identical anatomical location x, *i.e.*, $\mathbf{D_F}(x)$ and $\mathbf{D_M}(x)$, or with similar structural information is maximised. To this end, we introduce a Modality-Agnostic deep Structural Representation Network (MASR-Net, Sec. 3.2) and anatomy-aware contrastive learning paradigm (Sec. 3.3), followed by a multimodal similarity metric formulation with DNS (Sec. 3.4).

The overview of MASR-Net and anatomy-aware contrastive learning paradigm is illustrated in the lower and upper panels of Fig. 2, respectively. Our network first computes the image feature with an encoder-decoder network, extracts the deep structural information from the feature map with the DNS extractor, and encodes them with the feature squeezing module. The proposed network is trained with non-linear intensity transformation, followed by anatomy-aware contrastive learning. With these components, the complex structural and anatomical location-aware information are well reflected in the resulting deep intermediate structural representation.

3.2. Modality-Agnostic Deep Structural Representation Network

Feature extraction. We first leverage a feed-forward 3D convolutional neural network (CNN) to extract the image feature $\mathbf{h}$ for the input image I. The proposed CNN network is built with a 4-level encoder-decoder structure with skip connection [47], which is composed of 3D convolution layers, LeakyReLU activations [31] and uses BlurPool [61] and trilinear interpolation for downsampling and upsampling, respectively, resulting in a maximum striding factor of 8. The network takes the input image $I \in \mathbb{R}^{C \times H \times W \times D}$ and outputs an image feature map $\mathbf{h} \in \mathbb{R}^{C_h \times H \times W \times D}$. The details of the network are shown in supplementary material.
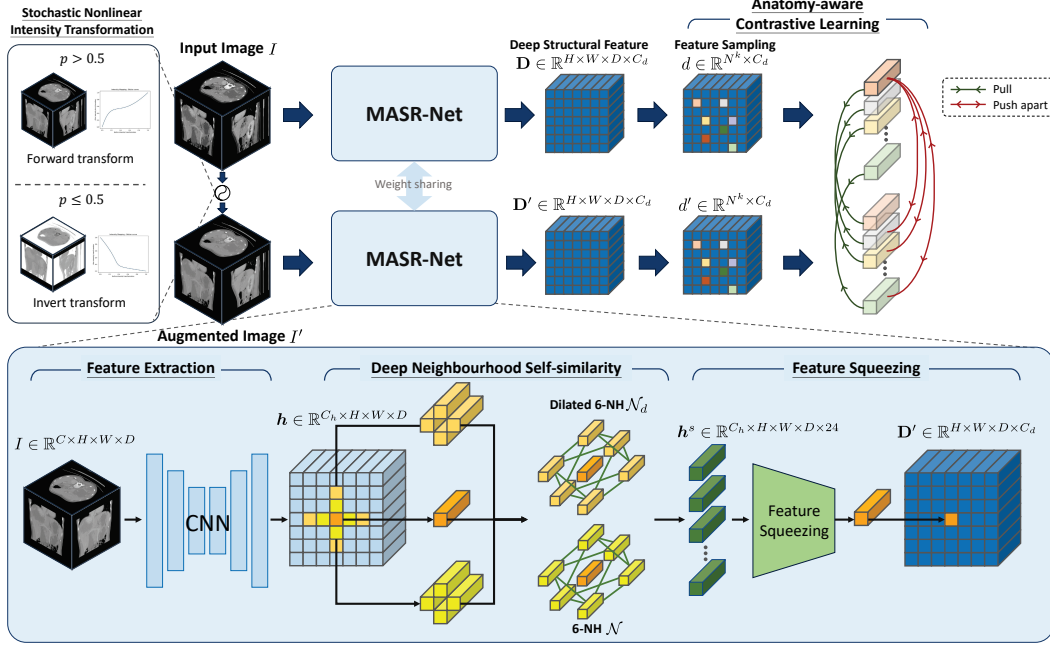
Figure 2. Overview of the Modality-Agnostic Deep Structural Representation Network (MASR-Net) and anatomy-aware contrastive learning paradigm. For brevity, we visualize the 3D feature maps in a 2D aspect. Only negative pairs of the first feature vector are shown.

Deep Neighbourhood Self-similarity (DNS). The vanilla local self-similarity descriptor [48] captures internal geometric layouts of local self-similarities within images by computing the pairwise distance between the patch center and its neighbourhood within a patch *at the pixel-level*. Yet, this formulation may be sensitive to the noise presented in the image, especially for medical images. In contrast, our proposed DNS is computed *at the feature level* and avoids using the patch center for the calculation, capable of capturing complex structural information of the image beyond the local neighbour context and less sensitive to the noise presented in medical images.

Formally, given the feature map $h \in \mathbb{R}^{C_h \times H \times W \times D}$ of input image I, a patch centred at $x \in \Omega$ and a certain neighbourhood layout $\mathcal{N}$, the proposed deep neighbourhood self-similarity $\mathbf{S}$ is given by:

$$\mathbf{S}(h, x, y) = \exp\left(-\sum_{y' \in \mathcal{N}(x)} \frac{(h(y) - h(y'))^2}{\sigma^2}\right), y' \neq y, \quad (1)$$

where $y, y' \in \mathcal{N}(x)$ defines the neighbour location of x. The denominator σ^2 is a noise estimator, defined as the mean of all patch distances, *i.e.*, $\frac{1}{|\mathcal{N}(x)|}\sum_{y' \in \mathcal{N}(x)}(h(y) - h(y'))^2$, where $|\mathcal{N}(x)|$ represents the number of neighbour voxels in $\mathcal{N}(x)$. We follow [20] to further restrict the pairwise distance calculations within the six-neighbourhood (6-NH) with a Euclidean distance of $\sqrt{2}$ between them, reducing the computation complexity of DNS, *i.e.*, reduce the number of unique pair-wise distances from 15 to 12.

To further maximize the discriminability of the computed feature, we compute two sets of DNS from h using two different neighbourhood layouts, *i.e.*, $\mathcal{N}$ and $\mathcal{N}_d$. We define $\mathcal{N}$ and $\mathcal{N}_d$ to be the direct 6-NH and dilated 6-NH layouts, respectively, as shown in Fig. 2. The DNS of the two neighbourhood layouts is then concatenated to form a 5-D feature map $h^s \in \mathbb{R}^{C_h \times H \times W \times D \times 24}$, containing the deep structural information of the feature map h.

Feature Squeezing. To compute a compact, intermediate DSIR from the 5-D DNS feature map h^s, we encode the high-dimensional 5-D feature map h^s into a compact 4D DNS descriptor using a feature squeezing module. The feature squeezing module consists of a single-layer linear layer with C_h perceptions, followed by a feed-forward convolution head. It first takes the 5-D DNS feature vector h^s as input and encodes it into a compact deep structural embedding $h^c \in \mathbb{R}^{H \times W \times D \times 24}$ using linear projection. The feed-forward convolution head is composed of two 3D convolution layers (kernel size $= 3^3$) with LeakyReLU activation in between the layers. It further encodes the compact deep structural embedding h^c to the DSIR $\mathbf{D} \in \mathbb{R}^{H \times W \times D \times C_d}$.

3.3. Anatomy-aware Contrastive Learning

The deep intermediate structural representation $\mathbf{D}$ using DNS is a regional descriptor that is able to capture local geometric structures in the feature map while suppressing appearance variation inside it. However, most regions in the image may share similar local geometric structures or suffer from image noise, causing ambiguity in matching the true anatomical correspondence. To further enhance the dis-

criminability of $\mathbf{D}$, we introduce an anatomy-aware contrastive learning paradigm that learns to maximize the similarity of the deep structural embeddings of the input image and augmented image at the identical anatomical locations while penalizing the similarity of that from different anatomical locations. Our anatomy-aware contrastive learning is depicted in the upper panel of Fig. 2.

Stochastic non-linear intensity transformation. To simulate the heterogeneous intensity distribution of the anatomical structures or lesions across multi-modal scans, we perform stochastic non-linear intensity augmentation to the augmented images, as shown in Fig. 2. Specifically, we employ the Bézier curve [59], which is controlled by $n + 1$ random control points $P_i \in \mathbb{R}^2$ uniformly sampled within $[0, 1]$ and a monotonic nonlinear transformation, to augment the intensity of the augmented image. The monotonicity ensures a one-to-one correspondence between pixel values, thereby maintaining the geometric structures of the image. Formally, the Bézier curve $B(t)$ is defined as:

$$B(t) = \sum_{i=0}^{n} P_i \cdot b_{i,n}(t), \tag{2}$$

where P_i are the control points, $b_{i,n}(t)$ are the Bernstein polynomials, n is the number of control points minus one, and t ranges from 0 to 1. Furthermore, to enforce the contrast invariant properties to the deep structural embeddings, we randomly invert the image intensity. The resulting nonlinear intensity transformation is then applied to the image intensity values as follows:

$$I'(x) = \begin{cases} B(I(x)), & \text{if } p > \delta \\ B(1 - I(x)), & \text{if } p \leq \delta \end{cases} \quad x \in \Omega, \tag{3}$$

where $p \in [0, 1]$ is the randomly sampled probability and δ is the threshold for the intensity inversion. The stochastic non-linear intensity transformation diversifies the intensity distribution of the structural information in the training set while preserving the crucial structural information between input and augmented image, improving the robustness of the learned deep intermediate structural representation across different modalities.

Contrastive loss. Let $\mathbf{D}, \mathbf{D}'$ denote the DSIR of the input image I and augmented image I', respectively. For optimal contrastive learning, the goal is to maximize the similarity of $\mathbf{D}$ and $\mathbf{D}'$ at identical anatomical locations while penalizing the similarity of that from different anatomical locations. Since I and I' are originated from the same image, $\mathbf{D}(x), \mathbf{D}'(x)$ are the deep structural embeddings sampled at identical anatomical locations given any $x \in \Omega$ sampled in the foreground of I.

For optimal contrastive learning, we randomly sample N^k feature vectors from $\mathbf{D}$ and $\mathbf{D}'$, respectively. Let $d_i = ||\mathbf{D}(i)||_2$ and $d_i' = ||\mathbf{D}'(i)||_2$ be the unit vector of $D(i)$ sampled with spatial indice i, where $i = 1, \ldots, N^k$. Indices

in correspondence d_i and d_i^+ are positive pairs. Similarly, d_i and d_j^-, where $j = 1, \ldots, N^k$ and $j \neq i$, are negative pairs. The contrastive loss is defined as:

$$\ell(d_i, d_i^+, d_j^-) = -\log \left[\frac{\exp(d_i \cdot d_i^+/\tau)}{\exp(d_i \cdot d_i^+/\tau) + \sum_{j=1, j \neq i}^{N^k} \exp(d_i \cdot d_j^-/\tau)} \right]. \tag{4}$$

3.4. Multimodal Similarity Using DNS

Without the loss of generality, given $\mathbf{D}_F$ and $\mathbf{D}_M$ are the DSIRs of F and M estimated by MASR-Net, the multimodal image registration problem with DNS can be formulated as:

$$\phi^* = \arg\min_\theta \mathcal{L}_{sim}(\psi_\sigma(\mathbf{D}_F), \psi_\sigma(\mathbf{D}_M(\phi_\theta))) + \lambda \mathcal{L}_{reg}(\phi_\theta), \tag{5}$$

where ϕ^* denotes the optimal displacement field, the ψ_σ is a Gaussian smoothing function with variance σ, $\mathcal{L}_{sim}(\mathbf{D}_F, \mathbf{D}_M(\phi_\theta))$ denotes the negated cosine similarity function that quantifies the dissimilarity between the $\mathbf{D}_F$ and the $\mathbf{D}_M(\phi_\theta)$ spatially transformed by ϕ, and $\mathcal{L}_{reg}(\phi) = ||\nabla\phi||_2^2$ represents the smoothness regularization function that penalizes implausible solution. Our proposed DNS is differentiable and can be adapted to iterative instance optimization methods [40, 51] and learning-based registration frameworks [6, 35–37]. To exemplify the flexibility of DSIR, we parameterize the registration function with one learning-based method, LapIRN [37], and one 3-level non-rigid instance optimization method (IO) [38], with implementation details shown in the supplementary material.

4. Experiments

4.1. Data and Pre-processing

Liver Multiphase CT. We collected a 3D multi-phase contrast-enhanced liver CT dataset from hospitals in China to evaluate the *inter-patient registration*. This dataset consists of 1966 patients with tumors, and each patient contains scans in three phases, *i.e.*, pre-contrast, arterial and venous phases, acquired from the CT scanner at different time points after the contrast agent injection. There are large non-linear deformations and heterogeneous image intensities across multi-phase CT due to gravity, muscle contractions, and contrast agent injection. We divided the dataset into 1886/10/70 cases for training, validation, and test sets. The liver and tumour masks of the validation and test sets are manually annotated and checked by a senior radiologist.

Abdomen MR-CT. We evaluated our method on the Abdomen MR-CT task of the Learn2Reg challenge 2021 [22]. The dataset comprises 8 sets of paired MR and CT scans and 90 unpaired CT/MR scans, both depicting the abdominal region of a single patient and presenting notable deformations. Since there are only limited paired MR and CT scans,

Table 1 (top part):

| Method | Metric | Pre-contrast ← Venous & Arterial | | | | | Arterial ← Venous & Pre-contrast | | | | |
| | | Tumour | | Organ | | $\%\|J_\phi\|<0\downarrow$ | Tumour | | Organ | | $\%\|J_\phi\|<0\downarrow$ |
		DSC ↑	HD95 ↓	DSC ↑	HD95 ↓		DSC ↑	HD95 ↓	DSC ↑	HD95 ↓	
Initial	–	75.51 ± 20.64	4.12 ± 4.00	88.03 ± 8.77	3.80 ± 3.31	–	77.84 ± 19.86	3.61 ± 3.19	90.15 ± 7.41	3.15 ± 2.68	–
ANTs	MI	76.52 ± 20.16	3.97 ± 3.15	87.40 ± 9.75	3.88 ± 2.91	0.00 ± 0.00	80.18 ± 18.47	3.42 ± 2.73	90.07 ± 8.04	3.22 ± 2.80	0.00 ± 0.00
NiftyReg	MI	77.10 ± 17.00	3.54 ± 2.54	90.87 ± 6.57	2.98 ± 2.10	0.35 ± 1.20	79.22 ± 16.54	3.18 ± 2.07	92.39 ± 5.61	2.53 ± 1.79	0.29 ± 1.18
DEEDs	MIND	79.65 ± 15.30	3.14 ± 1.98	93.90 ± 3.47	2.09 ± 1.50	0.23 ± 0.72	80.57 ± 15.23	3.42 ± 2.06	94.80 ± 2.75	$\mathbf{1.84} \pm 1.04$	0.12 ± 0.13
VM	NMI	75.78 ± 17.35	3.62 ± 2.69	91.70 ± 4.43	2.81 ± 1.89	0.00 ± 0.00	76.62 ± 18.19	3.23 ± 2.30	92.69 ± 3.81	2.45 ± 1.63	0.00 ± 0.00
VM	MIND	75.16 ± 17.07	3.94 ± 5.51	91.70 ± 4.92	2.77 ± 2.05	0.00 ± 0.00	75.18 ± 17.71	3.63 ± 5.11	92.15 ± 4.39	2.55 ± 1.65	0.00 ± 0.00
LapIRN	NMI	78.80 ± 15.37	3.33 ± 2.80	93.53 ± 3.75	2.35 ± 1.52	0.01 ± 0.02	80.17 ± 15.38	2.87 ± 2.00	94.48 ± 3.72	1.97 ± 1.53	0.02 ± 0.16
LapIRN	MIND	77.50 ± 16.71	3.96 ± 3.28	93.32 ± 4.41	2.46 ± 1.81	0.00 ± 0.01	79.49 ± 15.00	3.19 ± 2.44	94.29 ± 4.19	2.07 ± 1.77	0.01 ± 0.04
LapIRN (ours)	DNS	79.72 ± 14.44	3.06 ± 2.23	94.07 ± 3.36	$\mathbf{2.08} \pm 1.45$	0.00 ± 0.01	80.66 ± 14.56	$\mathbf{2.71} \pm 1.85$	94.73 ± 3.09	1.84 ± 1.25	0.02 ± 0.18
IO	MIND	76.27 ± 16.44	3.66 ± 2.74	92.54 ± 3.41	2.63 ± 1.32	0.08 ± 0.37	76.91 ± 16.14	3.54 ± 2.23	92.74 ± 3.45	2.70 ± 1.31	0.12 ± 0.51
IO (ours)	DNS	$\mathbf{80.43} \pm 13.72$	$\mathbf{2.94} \pm 2.23$	$\mathbf{94.26} \pm 3.32$	2.10 ± 1.50	0.03 ± 0.20	$\mathbf{81.07} \pm 13.83$	2.74 ± 1.80	$\mathbf{94.89} \pm 2.92$	1.85 ± 1.22	0.06 ± 0.35

Table 1 (bottom part):

| Method | Metric | Venous ← Arterial & Pre-contrast | | | | | Average Score across Three Tasks | | | | |
| | | Tumour | | Organ | | $\%\|J_\phi\|<0\downarrow$ | DSC ↑ | HD95 ↓ | $\%\|J_\phi\|<0\downarrow$ | T_{Test} | # Param |
		DSC ↑	HD95 ↓	DSC ↑	HD95 ↓						
Initial	–	78.10 ± 19.85	3.59 ± 3.07	88.96 ± 7.49	3.50 ± 2.41	–	83.10 ± 12.60	3.63 ± 2.97	–	–	–
ANTs	MI	81.14 ± 17.14	3.37 ± 2.69	89.20 ± 8.28	3.52 ± 2.46	0.00 ± 0.00	84.09 ± 12.84	3.56 ± 2.60	0.00 ± 0.00	$250.05 \pm 367.42*$	–
NiftyReg	MI	78.53 ± 17.05	3.32 ± 2.28	91.19 ± 6.80	2.90 ± 2.06	0.32 ± 1.19	84.88 ± 9.92	3.07 ± 1.98	0.32 ± 1.19	$79.17 \pm 30.77*$	–
DEEDs	MIND	81.28 ± 14.65	3.32 ± 1.87	$\mathbf{94.58} \pm 2.44$	$\mathbf{2.11} \pm 1.14$	0.11 ± 0.14	87.46 ± 8.97	2.65 ± 1.72	0.15 ± 0.43	$47.92 \pm 12.77*$	–
VM	NMI	77.20 ± 17.33	3.28 ± 2.09	92.65 ± 3.31	2.60 ± 1.43	0.00 ± 0.00	84.44 ± 9.47	3.00 ± 1.83	0.00 ± 0.00	0.14 ± 0.01	1.14M
VM	MIND	75.94 ± 17.18	3.34 ± 1.54	92.25 ± 4.07	2.63 ± 1.54	0.00 ± 0.00	83.73 ± 9.33	3.14 ± 2.57	0.00 ± 0.00	0.16 ± 0.01	1.14M
LapIRN	NMI	81.37 ± 14.05	2.86 ± 2.12	94.22 ± 3.46	2.24 ± 1.57	0.00 ± 0.00	87.10 ± 8.89	2.61 ± 1.91	0.00 ± 0.01	0.19 ± 0.01	1.59M
LapIRN	MIND	81.03 ± 14.56	3.09 ± 2.61	93.99 ± 3.98	2.33 ± 1.82	0.00 ± 0.00	86.61 ± 9.81	2.85 ± 2.29	0.00 ± 0.01	0.18 ± 0.01	1.59M
LapIRN (ours)	DNS	$\mathbf{81.62} \pm 13.42$	2.81 ± 1.99	94.37 ± 3.03	2.08 ± 1.41	0.00 ± 0.00	87.53 ± 8.65	2.43 ± 1.70	0.01 ± 0.01	0.18 ± 0.01	1.59M
IO	MIND	77.44 ± 16.26	3.48 ± 2.37	93.02 ± 3.30	2.56 ± 1.42	0.01 ± 0.02	84.82 ± 9.83	3.10 ± 1.90	0.07 ± 0.30	4.75 ± 0.44	-
IO (ours)	DNS	81.51 ± 13.67	$\mathbf{2.79} \pm 1.94$	94.53 ± 3.01	2.11 ± 1.48	0.00 ± 0.01	$\mathbf{87.78} \pm 8.41$	$\mathbf{2.42} \pm 1.70$	0.03 ± 0.19	5.05 ± 0.47	-

Table 1. Quantitative results on liver multi-phase CT registration task. $X \leftarrow Y$ represents the experiment of registering the CT scan in the Y phase to the CT scan in the X phase of the same patient. ↑: higher is better, and ↓: lower is better. Initial: initial results without registration. The registration runtime highlighted with an asterisk is reported in CPU time (in seconds).

we only evaluate the iterative optimization-based registration methods in this task. We use all 8 paired MR and CT scans as the test set. Each scan has four manually labeled anatomical structures: liver, spleen, right kidney, and left kidney. For MASR-Net, we split the 90 unpaired CT/MR scans into 85 and 5 scans for training and validation sets.

Brain MR T1w-T2w. We performed atlas-based registration on BraTS18 [†] and ISeg19 [‡] datasets [34, 53]. The training set consists of 135 cases, each with T1-weighted contrast (T1w) and T2-weighted contrast (T2w) images, and was resampled to 1mm^3 isotropic resolution. 10 cases contain full delineation of anatomical structures, including grey matter, white matter and cerebrospinal fluid (CSF) for both T1w and T2w modalities. We randomly select 1 case from the test set as the atlas and register the remaining 9 cases to the atlas. Similar to Abdomen MR-CT, we only evaluate iterative optimization-based registration methods on this task due to the limited training data.

4.2. Implementation

All DLIR methods, *i.e.*, DNS, Voxelmorph, and LapIRN, are developed and trained using Pytorch. All the meth-

[†] https://www.med.upenn.edu/sbia/brats2018.html
[‡] https://iseg2019.web.unc.edu/

| Method | Metric | Abdomen MR ← CT | | Brain MR T1w ↔ T2w | | |
		DSC ↑	T_{Test}	$DSC_{T1\leftarrow T2}$ ↑	$DSC_{T2\leftarrow T1}$ ↑	T_{Test}
Initial	–	37.32 ± 17.23	–	53.90 ± 0.70	53.90 ± 0.70	–
DEEDs	MIND	83.53 ± 8.58	$165.21 \pm 25.14*$	61.47 ± 0.96	60.70 ± 0.83	14.45 ± 0.54
ANTs	MI	39.43 ± 17.67	$38.54 \pm 3.51*$	61.00 ± 1.00	52.80 ± 1.10	18.46 ± 1.33
NiftyReg	MI	48.60 ± 31.60	$37.88 \pm 10.95*$	63.90 ± 1.10	61.90 ± 0.70	34.75 ± 3.04
IO (ours)	DNS	85.61 ± 5.95	4.52 ± 0.50	62.66 ± 0.73	61.91 ± 0.66	5.84 ± 0.54

Table 2. Quantitative results on Abdomen MR-CT and brain MR T1w-T2w registration tasks. ↑: higher is better, and ↓: lower is better. Initial: initial results in native space without registration.

ods are trained or executed on a standalone workstation equipped with 8 Nvidia V100 GPUs and an Intel Xeon Platinum 8163 CPU. We adopt the Adam optimizer [24] with a fixed learning rate of $1e^{-4}$ and batch size set to 1 for all learning-based approaches. For the training of MASR-Net, we set the number of sampling feature vectors N^k, temperature τ, the number of random control points n, threshold δ and C_d to 8196, 0.07, 3, 0.5 and 24, respectively. For the IO method, we adopt the 3-level multiresolution optimization strategy and Adam optimizer with a fixed learning rate of $1e^{-3}$ for each level.

4.3. Evaluation metrics and Baselines

To quantify the registration performance, we register each pair, propagate the anatomical segmentation map using the resulting transformation, and measure the region of interest
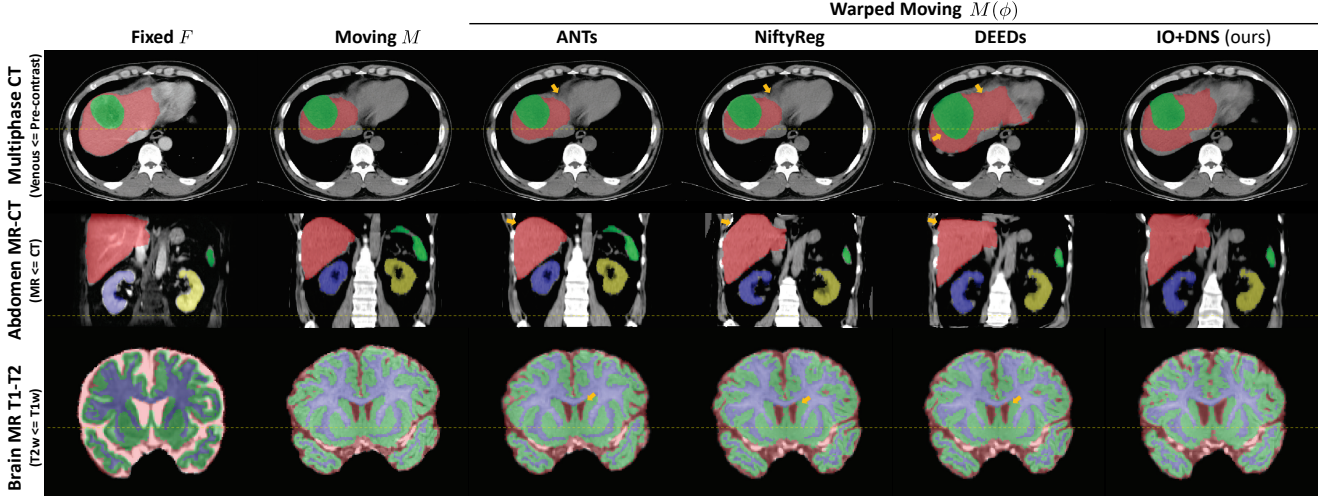
Figure 3. Example slices of resulting warped images using different registration methods on Liver multiphase CT, Abdomen MR-CT, and Brain MR T1w-T2w registration tasks. The warped anatomical segmentations are overlayed on the resulting images. Major registration artefacts are highlighted with yellow arrows. Multiphase CT: tumour (green) and liver (red). Abdomen MR-CT: spleen (green), liver (red), right kidney (yellow), and left kidney (blue). MR T1w-T2w: grey matter (green), white matter (blue) and cerebrospinal fluid (red).

(ROI) segmentation overlap using the Dice similarity coefficient (DSC). We also measure the 95% percentile of the Hausdorff distance (HD95) to represent the registration accuracy to the boundary of the ROI structures. As a negative determinant of the Jacobian at a voxel indicates local folding [2], we further compute the percentage of foldings ($\%|J_\phi| < 0$) within the deformation field to evaluate the plausibility of the deformation fields.

We compare our method with three state-of-the-art conventional registration methods (ANTs [4], NiftyReg [52] and DEEDs [18]) and two learning-based approaches (Voxelmorph [6] and LapIRN [41]). Specifically, we use the ANTs registration implementation in the publicly available ANTs software package [3]. Both ANTs and NiftyReg methods use a four-level multi-resolution strategy with adaptive gradient descent optimization and MI as the similarity measure. For Voxelmorph and LapIRN, we use official implementations and the best hyperparameter reported in their paper. By default, all learning-based methods are trained with two similarity metrics (NMI and MIND).

4.4. Results

Intra-subject Liver multiphase CT registration. Table 1 shows the results of the Liver multiphase CT registration. The initial DSC of the registration of venous and arterial CT to pre-contrast (Pre-contrast ← Venous & Arterial) is the lowest among all directions, implying that there is a relatively large misalignment between pre-contrast and enhanced CT scans. Overall, our proposed method (IO+DNS and LapIRN+DNS) achieves the best registration accuracy in terms of DSC and HD95 over all three registration directions, suggesting that our method is robust and accurate in

Methods	DSC ↑	HD95 ↓
Initial	81.77 ± 16.96	3.96 ± 3.66
Backbone network feature (Random initialization)	75.38 ± 21.03	5.90 ± 5.46
+Deep Neighbourhood Self-similarity	86.50 ± 13.43 (+11.12)	2.79 ± 2.30 (-3.11)
+Nonlinear Intensity & Contrastive Learning	87.13 ± 12.06 (+0.63)	2.57 ± 1.95 (-0.22)
+Gaussian Smoothing	87.35 ± 12.14 (+0.22)	2.52 ± 1.95 (-0.05)

Table 3. Influence of the deep neighbourhood self-similarity and contrastive learning to the registration model.

multiphase CT registration, especially in the alignment of tumor region, as shown in the first row of Fig. 3.

Flexibility and versatility. Compared to the results of LapIRN with different metrics, LapIRN with our DNS achieves the best overall registration performance in tumor and organ alignment, showing +2.22 and +0.75 DSC gains over the one with MIND in tumor and organ alignment, respectively. More importantly, when comparing the results of IO with DNS and MIND, there are significant improvements in registration accuracy of both tumor and organ (+4.16 and +1.72 gains in DSC). The result proves that our method is more robust and flexible with diverse registration frameworks than MIND. When using DNS with a simple iterative gradient optimization, the DSIR of DNS is more discriminative than MIND and can lead to less ambiguous matching during the iterative optimization. Besides, although NMI shows comparable DSC in multiphase CT registration, the qualitative results in Fig. 3 demonstrate that methods using MI/NMI fail to recover the large nonlinear misalignment in MR-CT registration.

Intra-subject abdomen MR-CT registration. Table 2 depicts the result of abdomen MR-CT registration.
Robust to large deformation. The initial DSC of this task is relatively low (37.32), indicating there are possible large spatial misalignments between image pairs. All conven-
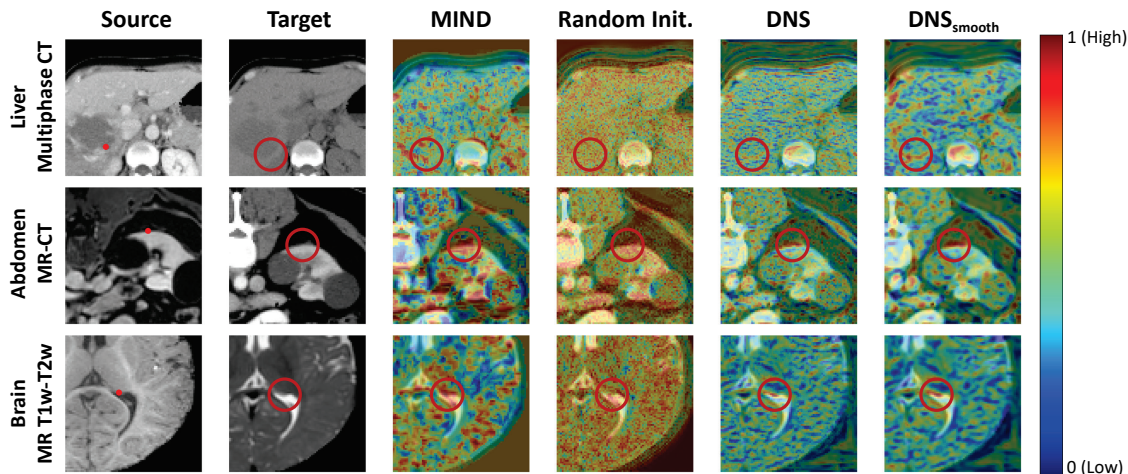
Figure 4. Visualization of the feature similarity maps across different modalities and anatomies. Random Init.: DNS without training. DNS$_{smooth}$: DNS with Gaussian smoothing. Each heatmap shows the similarity of the marked point (red pot) on the source image to every point in the target image. The feature extracted with our method (DNS and DNS$_{smooth}$) shows high discriminability at the boundary of liver tumour (1st row), kidney (2nd row) and anatomical structure of brain MR with large intensity variation (3rd row).

tional methods (ANTs and NiftyReg) with MI fail spectacularly in this task, suggesting that MI cannot register multimodal images with large deformation. In contrast to MI, structural image representation-based methods (DEEDs and DNS) achieve decent results in this task (83.53 and 85.61 DSC, respectively). It is worth noting that our DNS with a solely simple gradient descent-based optimization strategy (IO) outperforms DEEDs, which employ a more sophisticated convex optimization.

Inter-subject brain MR T1w-T2w registration. Table 2 shows the results of brain MR T1w-T2w registration. *Contrast invariance.* Brain MR T1w and T2w scans show a huge heterogeneous intensity distribution between anatomical structures across T1w and T2w modalities. Our method (IO+DNS) achieves on-par registration accuracy with conventional methods (DEEDs, ANTs and NiftReg), boosting the initial DSC from 53.90, 53.90 to 62.66 and 61.91, respectively, indicating that DNS is robust to contrast change in multimodal images. The feature similarity of DNS, as shown in Fig. 4, further suggests that DNS has high discriminability and is robust to images with heterogeneous intensity distribution.

4.4.1 Ablation Studies

Discriminability of the structural image representations. As demonstrated in Fig. 4, each heatmap shows the similarity of the marked point on the source image to every point in the target image on three multi-modal registration tasks. Our method produces less ambiguous heatmaps than MIND on tumor boundaries (1st row) and anatomical structures (2nd and 3rd rows), indicating our method is capable of capturing expressive and discriminative DSIR.
Effect of Deep Neighbourhood Self-similarity and con-

trastive learning. Tab. 3 demonstrates performance metrics for different registration models on the multiphase CT registration task (Pre-contrast ← Venous & Arterial). Although a recent study [13] reported that random features from CNN could be expressive, the DSIR of random initialize CNN has lower DSC than that of initial, suggesting that random initialization alone is insufficient for multi-modal registration. Moreover, there is a significant performance gain when the DNS module is added to the randomly initialized CNN (+11.12 DSC), suggesting the effectiveness of DNS in extracting deep structural information. Comparing the feature similarity analysis of random initialization (MASR-Net without training) and DNS in Fig. 4, one can observe that the proposed anatomy-aware contrastive learning effectively improves the discriminability of the DSIR of DNS, resulting in high feature similarity centralized in the true anatomical correspondence and relatively low feature similarity at the other anatomical structures.

5. Conclusion

This paper proposes a deep structural image representation learning method dedicated to multi-modality medical image registration. Our method leverages deep neighbourhood self-similarity to learn highly discriminative, contrast invariance structural representations for multimodal images. Anatomy-aware contrastive learning is introduced to further enhance the expressiveness and discriminability of the learned structural representation, reducing the ambiguity in matching the true anatomical correspondence between multimodal images. Comprehensive experiments demonstrate that our method achieves state-of-the-art registration accuracy compared with conventional local structural representations without the need for a dedicated optimizer.

References

[1] Fakhre Alam, Sami Ur Rahman, Sehat Ullah, and Kamal Gulati. Medical image registration in image guided surgery: Issues, challenges and research opportunities. *Biocybernetics and Biomedical Engineering*, 38(1):71–89, 2018. 1

[2] John Ashburner. A fast diffeomorphic image registration algorithm. *NeuroImage*, 38(1):95–113, 2007. 7

[3] Brian B Avants, Nick Tustison, and Gang Song. Advanced normalization tools (ANTS). *Insight j*, 2(365):1–35, 2009. 2, 7

[4] Brian B Avants, Nicholas J Tustison, Gang Song, Philip A Cook, Arno Klein, and James C Gee. A reproducible evaluation of ants similarity metric performance in brain image registration. *NeuroImage*, 54(3):2033–2044, 2011. 2, 7

[5] Xiaoyu Bai, Fan Bai, Xiaofei Huo, Jia Ge, Tony C. W. Mok, Zi Li, Minfeng Xu, Jingren Zhou, Le Lu, Dakai Jin, Xianghua Ye, Jingjing Lu, and Ke Yan. Matching in the wild: Learning anatomical embeddings for multi-modality images. *CoRR*, abs/2307.03535, 2023. 3

[6] Guha Balakrishnan, Amy Zhao, Mert R Sabuncu, John Guttag, and Adrian V Dalca. Voxelmorph: A learning framework for deformable medical image registration. *IEEE Transactions on Medical Imaging*, 38(8):1788–1800, 2019. 5, 7

[7] Alexander Bigalke, Lasse Hansen, Tony CW Mok, and Mattias P Heinrich. Unsupervised 3d registration through optimization-guided cyclical self-training. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 677–687. Springer, 2023. 2

[8] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 3

[9] Neel Dey, Jo Schlemper, Seyed Sadegh Mohseni Salehi, Bo Zhou, Guido Gerig, and Michal Sofka. Contrareg: Contrastive learning of multi-modality unsupervised deformable image registration. In *Medical Image Computing and Computer Assisted Intervention*, pages 66–77, 2022. 3

[10] Matthias Eisenmann et al. Biomedical image analysis competitions: The state of current participation practice. *arXiv preprint arXiv:2212.08568*, 2022. 3

[11] Fenja Falta, Lasse Hansen, and Mattias P Heinrich. Learning iterative optimisation for deformable image registration of lung ct with recurrent convolutional networks. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part VI*, pages 301–309. Springer, 2022. 2

[12] Xin Fan, Zi Li, Ziyang Li, Xiaolin Wang, Risheng Liu, Zhongxuan Luo, and Hao Huang. Automated learning for deformable medical image registration by jointly optimizing network architectures and objective functions. *IEEE Transactions on Image Processing*, 32:4880–4892, 2023. 3

[13] Jonathan Frankle, David J Schwab, and Ari S Morcos. Training batchnorm and only batchnorm: On the expressive power of random features in cnns. *arXiv preprint arXiv:2003.00152*, 2020. 8

[14] Benjamin Gutierrez-Becker, Diana Mateus, Loic Peter, and Nassir Navab. Guiding multimodal registration with learned optimization updates. *Medical image analysis*, 41:2–17, 2017. 2, 3

[15] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020. 3

[16] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. Momentum contrast for unsupervised visual representation learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9726–9735, 2020. 3

[17] Mattias P. Heinrich, Mark Jenkinson, Manav Bhushan, Tahreema N. Matin, Fergus Gleeson, Michael Brady, and Julia A. Schnabel. MIND: modality independent neighbourhood descriptor for multi-modal deformable registration. *Medical Image Analysis*, 16(7):1423–1435, 2012. 2, 3

[18] Mattias P. Heinrich, Mark Jenkinson, Michael Brady, and Julia A. Schnabel. Globally optimal deformable registration on a minimum spanning tree using dense displacement sampling. In *Medical Image Computing and Computer Assisted Intervention*, pages 115–122, 2012. 3, 7

[19] Mattias P Heinrich, Mark Jenkinson, Sir Michael Brady, and Julia A Schnabel. Globally optimal deformable registration on a minimum spanning tree using dense displacement sampling. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2012: 15th International Conference, Nice, France, October 1-5, 2012, Proceedings, Part III 15*, pages 115–122. Springer, 2012. 3

[20] Mattias P Heinrich, Mark Jenkinson, Bartlomiej W Papież, Sir Michael Brady, and Julia A Schnabel. Towards realtime multimodal fusion for image-guided interventions using self-similarities. In *Medical Image Computing and Computer-Assisted Intervention*, pages 187–194, 2013. 3, 4

[21] Olivier Henaff. Data-efficient image recognition with contrastive predictive coding. In *International conference on machine learning*, pages 4182–4192. PMLR, 2020. 3

[22] Alessa Hering, Lasse Hansen, Tony CW Mok, Albert CS Chung, Hanna Siebert, Stephanie Häger, Annkristin Lange, Sven Kuckertz, Stefan Heldmann, et al. Learn2reg: comprehensive multi-task medical image registration challenge, dataset and evaluation in the era of deep learning. *IEEE Transactions on Medical Imaging*, 2022. 3, 5

[23] Andrew Hoopes, Malte Hoffmann, Bruce Fischl, John Guttag, and Adrian V Dalca. Hypermorph: Amortized hyperparameter learning for image registration. In *International Conference on Information Processing in Medical Imaging*, 2021. 2

[24] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[25] Daewon Lee, Matthias Hofmann, Florian Steinke, Yasemin Altun, Nathan D Cahill, and Bernhard Scholkopf. Learning similarity measure for multi-modal 3d image registration.

In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 186–193. IEEE, 2009. 2, 3

[26] Zi Li, Lin Tian, Tony C. W. Mok, Xiaoyu Bai, Puyang Wang, Jia Ge, Jingren Zhou, Le Lu, Xianghua Ye, Ke Yan, and Dakai Jin. Samconvex: Fast discrete optimization for CT registration using self-supervised anatomical embedding and correlation pyramid. In *Medical Image Computing and Computer Assisted Intervention*, pages 559–569, 2023. 3

[27] Fengze Liu, Jinzheng Cai, Yuankai Huo, Chi-Tung Cheng, Ashwin Raju, Dakai Jin, Jing Xiao, Alan L. Yuille, Le Lu, Chien-Hung Liao, and Adam P. Harrison. JSSR: A joint synthesis, segmentation, and registration system for 3d multi-modal image alignment of large-scale pathological CT scans. In *Computer Vision - ECCV 2020 - 16th European Conference*, pages 257–274, 2020. 3

[28] Risheng Liu, Zi Li, Yuxi Zhang, Xin Fan, and Zhongxuan Luo. Bi-level probabilistic feature learning for deformable image registration. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI*, pages 723–730. ijcai.org, 2020. 2

[29] Risheng Liu, Zi Li, Xin Fan, Chenying Zhao, Hao Huang, and Zhongxuan Luo. Learning deformable image registration from optimization: Perspective, modules, bilevel training and beyond. *IEEE Transactions on Pattern Analysis Machine Intelligence*, 44(11):7688–7704, 2022. 2, 3

[30] Chao Lu, Sudhakar Chelikani, Xenophon Papademetris, Jonathan P Knisely, Michael F Milosevic, Zhe Chen, David A Jaffray, Lawrence H Staib, and James S Duncan. An integrated approach to segmentation and nonrigid registration for application in image-guided pelvic radiotherapy. *Medical Image Analysis*, 15(5):772–785, 2011. 1

[31] Andrew L Maas, Awni Y Hannun, Andrew Y Ng, et al. Rectifier nonlinearities improve neural network acoustic models. In *Proc. icml*, page 3. Atlanta, GA, 2013. 3

[32] Frederik Maes, Andre Collignon, Dirk Vandermeulen, Guy Marchal, and Paul Suetens. Multimodality image registration by maximization of mutual information. *IEEE transactions on Medical Imaging*, 16(2):187–198, 1997. 2

[33] Calvin R Maurer and J Michael Fitzpatrick. A review of medical image registration. *Interactive image-guided neurosurgery*, 1:17–44, 1993. 1

[34] Bjoern H. Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, et al. The multimodal brain tumor image segmentation benchmark (brats). *IEEE Transactions on Medical Imaging*, 34(10):1993–2024, 2015. 6

[35] Tony CW Mok and Albert Chung. Fast symmetric diffeomorphic image registration with convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4644–4653, 2020. 2, 5

[36] Tony CW Mok and Albert Chung. Affine medical image registration with coarse-to-fine vision transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20835–20844, 2022.

[37] Tony CW Mok and Albert CS Chung. Conditional deformable image registration with convolutional neural 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*, pages 35–45. Springer, 2021. 5

[38] Tony CW Mok and Albert CS Chung. Robust image registration with absent correspondences in pre-operative and follow-up brain mri scans of diffuse glioma patients. In *International MICCAI Brainlesion Workshop*, pages 231–240. Springer, 2022. 5

[39] Tony CW Mok and Albert CS Chung. Unsupervised deformable image registration with absent correspondences in pre-operative and post-recurrence brain tumor mri scans. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 25–35. Springer, 2022. 1, 2

[40] Tony CW Mok, Zi Li, Yingda Xia, Jiawen Yao, Ling Zhang, Jingren Zhou, and Le Lu. Deformable medical image registration under distribution shifts with neural instance optimization. In *International Workshop on Machine Learning in Medical Imaging*, pages 126–136. Springer, 2023. 5

[41] Tony C. W. Mok and Albert C. S. Chung. Large deformation diffeomorphic image registration with laplacian pyramid networks. In *Medical Image Computing and Computer Assisted Intervention*, pages 211–221, 2020. 7

[42] Markus Müller, Linn Emilie Sævil Helljesen, Raphael Prevost, Ivan Viola, Kim Nylund, Odd Helge Gilja, Nassir Navab, and Wolfgang Wein. Deriving anatomical context from 4d ultrasound. In *Vcbm*, pages 173–180, 2014. 3

[43] Taesung Park, Alexei A Efros, Richard Zhang, and Jun-Yan Zhu. Contrastive learning for unpaired image-to-image translation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16*, pages 319–345. Springer, 2020. 3

[44] Nicolas Pielawski, Elisabeth Wetzer, Johan Öfverstedt, Jiahao Lu, Carolina Wählby, Joakim Lindblad, and Natasa Sladoje. Comir: Contrastive multimodal image representation for registration. *Advances in neural information processing systems*, 33:18433–18444, 2020. 2, 3

[45] Josien PW Pluim, JB Antoine Maintz, and Max A Viergever. Image registration by maximization of combined mutual information and gradient information. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2000: Third International Conference, Pittsburgh, PA, USA, October 11-14, 2000. Proceedings 3*, pages 452–461. Springer, 2000. 2, 3

[46] Michael JD Powell et al. The bobyqa algorithm for bound constrained optimization without derivatives. *Cambridge NA Report NA2009/06, University of Cambridge, Cambridge*, 26, 2009. 3

[47] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015. 3

[48] Eli Shechtman and Michal Irani. Matching local self-similarities across images and videos. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2007. 2, 4

[49] Dinggang Shen and Christos Davatzikos. Hammer: hierarchical attribute matching mechanism for elastic registration. *IEEE Transactions on medical imaging*, 21(11):1421–1439, 2002. 3

[50] Yucheng Shu, Hao Wang, Bin Xiao, Xiuli Bi, and Weisheng Li. Medical image registration based on uncoupled learning and accumulative enhancement. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part IV*, pages 3–13. Springer, 2021. 2

[51] Hanna Siebert, Lasse Hansen, and Mattias P Heinrich. Fast 3d registration with accurate optimisation and little learning for learn2reg 2021. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 174–179. Springer, 2021. 5

[52] Wei Sun, Wiro J Niessen, and Stefan Klein. Free-form deformation using lower-order b-spline for nonrigid image registration. In *Medical Image Computing and Computer Assisted Intervention*, pages 194–201, 2014. 2, 7

[53] Yue Sun, Kun Gao, Zhengwang Wu, Guannan Li, Xiaopeng Zong, Zhihao Lei, Ying Wei, Jun Ma, Xiaoping Yang, Xue Feng, Li Zhao, Trung Le Phan, Jitae Shin, Tao Zhong, Yu Zhang, Lequan Yu, Caizi Li, Ramesh Basnet, M. Omair Ahmad, M. N. S. Swamy, Wenao Ma, Qi Dou, Toan Duc Bui, Camilo Bermudez Noguera, Bennett A. Landman, Ian H. Gotlib, Kathryn L. Humphreys, Sarah Shultz, Longchuan Li, Sijie Niu, Weili Lin, Valerie Jewells, Dinggang Shen, Gang Li, and Li Wang. Multi-site infant brain segmentation algorithms: The iseg-2019 challenge. *IEEE Trans. Medical Imaging*, 40(5):1363–1376, 2021. 6

[54] Aäron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *CoRR*, abs/1807.03748, 2018. 3

[55] Paul Viola and William M Wells III. Alignment by maximization of mutual information. *International journal of computer vision*, 24(2):137–154, 1997. 2

[56] Wolfgang Wein, Shelby Brunke, Ali Khamene, Matthew R Callstrom, and Nassir Navab. Automatic ct-ultrasound registration for diagnostic imaging and image-guided intervention. *Medical image analysis*, 12(5):577–585, 2008. 3

[57] Yiming Xiao, Hassan Rivaz, Matthieu Chabanas, Maryse Fortin, Ines Machado, Yangming Ou, Mattias P Heinrich, Julia A Schnabel, Xia Zhong, Andreas Maier, et al. Evaluation of mri to ultrasound registration methods for brain shift correction: the curious2018 challenge. *IEEE transactions on medical imaging*, 39(3):777–786, 2019. 1

[58] Ke Yan, Jinzheng Cai, Dakai Jin, Shun Miao, Dazhou Guo, Adam P. Harrison, Youbao Tang, Jing Xiao, Jingjing Lu, and Le Lu. SAM: self-supervised learning of pixel-wise anatomical embeddings in radiological images. *IEEE Trans. Medical Imaging*, 41(10):2658–2669, 2022. 3

[59] Mohammad Asif Zaman and Shuvro Chowdhury. Modified bézier curves with shape-preserving characteristics using differential evolution optimization algorithm. *Advances in Numerical Analysis*, 2013:1–8, 2013. 5

[60] Qi Zeng, Shahed Mohammed, Emily HT Pang, Caitlin Schneider, Mohammad Honarvar, Julio Lobo, Changhong Hu, James Jago, Gary Ng, Robert Rohling, et al. Learning-based us-mr liver image registration with spatial priors. In *International conference on medical image computing and computer-assisted intervention*, pages 174–184. Springer, 2022. 3

[61] Richard Zhang. Making convolutional networks shift-invariant again. In *International conference on machine learning*, pages 7324–7334. PMLR, 2019. 3