

Unsupervised multimodal brain MRI registration with complementary learned and structural descriptors.

DM BAPPY

Korea University

Hyeongboo Baek

hbbaek359@gmail.com

University of Seoul

Research Article

Keywords: Image registration, Multimodal MRI, Contrastive learning, Diffeomorphism, Unsupervised learning, Brain

Posted Date: October 23rd, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11462321/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Additional Declarations: The authors declare no competing interests.

Unsupervised multimodal brain MRI registration with complementary learned and structural descriptors

D M Bappy¹ and Hyeongboo Baek^{2,*}

¹College of Medicine, Korea University, Seoul, Republic of Korea

²Department of Artificial Intelligence, University of Seoul, Seoul, Republic of Korea

*Correspondence: hgbaek359@gmail.com

D M Bappy: dewan8511@gmail.com

Abstract

Deformable registration of T1- and T2-weighted brain magnetic resonance images is difficult because tissue intensities differ across contrasts. We investigate whether learned and hand-crafted descriptors provide complementary similarity information for unsupervised registration. A Cross-Modal Descriptor Network learns contrast-invariant voxel embeddings through self-supervised contrastive training on aligned same-subject pairs. These embeddings are combined with the modality-independent neighbourhood descriptor with self-similarity context (MIND-SSC) to train a compact diffeomorphic attention U-Net without anatomical labels. In inter-subject T1-to-T2 registration on the IXI dataset ($n = 45$), the hybrid achieved a mean Dice coefficient of 0.700 and a 95th-percentile Hausdorff distance of 3.21 mm, with no observed folded voxels. It outperformed the evaluated classical, learned, structural-descriptor, transformer, and foundation-model baselines ($p < 0.05$, paired Wilcoxon tests). Gains were largest for small deep gray-matter structures, and the hybrid also performed best in the evaluated unseen-scanner and BraTS glioma experiments. Descriptors are used only during training; inference requires one 83 ms forward pass of a 0.247 M-parameter network. These results support combining learned and structural descriptors for accurate and computationally efficient multimodal brain registration.

Keywords: Image registration; Multimodal MRI; Contrastive learning; Diffeomorphism; Unsupervised learning; Brain

1 Introduction

Multi-contrast magnetic resonance imaging (MRI) is a cornerstone of clinical and research neuroimaging. T1-weighted (T1w) scans give high gray–white matter contrast and are the standard input for morphometry, whereas T2-weighted (T2w) scans emphasize fluid and many pathologies, including edema, demyelination, and tumor infiltration. Radiologists routinely read the two contrasts side by side, and automated pipelines increasingly combine them for segmentation, lesion detection, surgical and radiotherapy planning, and longitudinal monitoring. All of these uses depend on accurate spatial correspondence between images. When the contrasts come from different sessions, different subjects, or a reference atlas, that correspondence must be established by deformable registration, and every downstream measurement inherits its errors.

Multimodal registration is harder than its single-contrast counterpart for a structural reason. The mapping between T1w and T2w intensities is *non-monotonic*: cerebrospinal fluid (CSF) is dark in T1w but bright in T2w, while white matter shows the reverse ordering. No monotone intensity transform can make the two images directly comparable, so mean-squared error and cross-correlation, the standard objectives for unsupervised single-modality registration, are uninformative or misleading across contrasts. Classical multimodal registration therefore relied on statistical criteria such as mutual information [1, 2] and on modality-independent structural descriptors such as MIND [3]. These criteria work well but are usually optimized iteratively for each image pair, which takes minutes and must be tuned per application.

Learning-based registration [4, 5] replaces per-pair optimization with a network that predicts a dense deformation in a single forward pass, cutting registration time from minutes to milliseconds. Most of these methods, however, target the *unimodal* setting, where local normalized cross-correlation (NCC) is an adequate training loss. When the fixed and moving images have different contrasts and no ground-truth deformation is available, a central question remains open: *what similarity should the network be trained to maximize?* Recent studies learn contrast-invariant representations for registration [6, 7, 8], and large foundation models [9, 10] aim for task-agnostic generality. What has been missing is a controlled, resolution-matched comparison that relates a modern *learned* descriptor to a strong *hand-crafted* one, and that asks whether the two carry the same information or complementary information.

A second, often overlooked issue is the *validity* of the deformation. Overlap metrics such as Dice can be raised by transforms that fold space, that is, transforms whose Jacobian determinant is non-positive in places. Folded transforms are not invertible, do not preserve topology, and are unsuitable for voxel-based morphometry or for propagating contours in treatment planning. A useful multimodal method should therefore be judged on accuracy and on deformation regularity together.

This paper treats the cross-modal similarity objective as the main object of study. We argue that, for unsupervised multimodal registration, the choice of similarity matters more than the choice of backbone, and that learned and hand-crafted descriptors are complementary rather than competing. Combined in a compact diffeomorphic network, they give the most accurate results in our comparison with a guaranteed invertible transform, at a small fraction of the computational cost of recent large models and without anatomical labels. Our contributions are:

- A **self-supervised Cross-Modal Descriptor Network (CMDN)** that learns a dense, per-voxel, contrast-invariant embedding through a noise-contrastive objective on aligned same-subject T1w/T2w pairs, with no anatomical labels. A random deformation shared by both contrasts creates dense positive correspondences, and contrastive negatives prevent representation collapse.
- A **hybrid similarity** that combines the learned descriptor with the hand-crafted MIND-SSC descriptor, together with evidence that the two are complementary: the hybrid significantly outperforms either one alone, both on average and on most individual subjects.

- A **resolution-matched evaluation** against classical, single-modality, hand-crafted, recent transformer, and foundation-model baselines that reports overlap, boundary error, and deformation regularity side by side, and that tests generalization to an unseen scanner and to brains with glioma.

The remainder of the paper is organized as follows. Section 2 reviews related work. Section 3 describes the registration backbone, the CMDN, and the hybrid objective. Section 4 details the data, baselines, and evaluation protocol. Section 5 presents quantitative and qualitative results, ablations, generalization experiments, and efficiency. Section 6 discusses the findings, their clinical implications, and limitations, and Section 7 concludes.

2 Related work

Deformable medical image registration has a long history in both classical optimization and deep learning [11, 12]. We organize prior work by how the cross-modal similarity is defined, since this is the axis most relevant to this paper.

2.1 Classical registration and similarity measures

Early deformable methods optimized a parametric or non-parametric transform against an intensity criterion. Examples include free-form deformations based on B-splines [13], large-deformation diffeomorphic metric mapping (LDDMM) [14], diffeomorphic demons [15], symmetric normalization (SyN) [16], and GPU-accelerated free-form deformation [17]. For multimodal data, intensity differences motivated statistical criteria, most notably mutual information [1, 2] and its self-similarity-weighted variants [18]. A complementary line developed *modality-independent structural descriptors*. MIND and its self-similarity-context extension MIND-SSC [3, 19] encode each voxel by the pattern of similarities among nearby patches. Because this pattern depends on local geometry rather than on absolute intensity, it is largely unaffected by the contrast-specific intensity mapping. Such descriptors are robust and interpretable, but they are fixed rather than adapted to the data, and they are typically used inside slow per-pair optimization.

2.2 Learning-based registration

VoxelMorph [20, 4] established the paradigm of a U-Net combined with a spatial transformer and trained without supervision, giving amortized inference. De Vos *et al.* [21] proposed unsupervised affine and deformable networks at about the same time. Diffeomorphic variants integrate a stationary velocity field by scaling and squaring [22] or use Laplacian pyramids for large deformations [23], and cycle consistency has been used to improve regularity [24]. A more recent wave enlarges the effective receptive field or refines estimates from coarse to fine: transformer backbones (TransMorph [5], ModeT [25], NestedMorph [26]), frequency-domain models (Fourier-Net [27]), correlation-aware multilayer perceptrons (CorrMLP [28]), recursive pyramids (RDP [29]), and correlation-mining CNN–transformer hybrids (CorrMorph [30]). These architectural advances are largely independent of the similarity question and are mostly evaluated in the unimodal setting with a cross-correlation loss. As our experiments show, when such models are applied to multimodal data with their native objective, they reach usable overlap only through irregular, non-diffeomorphic deformations.

The Learn2Reg challenge [31] has helped standardize evaluation across many of these methods.

2.3 Multimodal-specific learning

Another line of work addresses the cross-modal objective directly. Image-to-image translation reduces the problem to unimodal registration [32], adversarial training learns a modality-agnostic criterion [33], and disentangled representations separate anatomical content from appearance [34]. Closest to our work, contrastive and structural representation learning yields contrast-invariant features for registration: CoMIR [6] learns rotation-equivariant shared representations, ContraReg [7] uses a patch-wise contrastive loss inside the registration objective, and MASR [8] learns modality-agnostic structural representations. CMDN belongs to this family and draws on the InfoNCE objective from self-supervised representation learning [35, 36]. Unlike prior work in this family, we measure explicitly how a learned descriptor relates to a hand-crafted one and show that their combination outperforms each.

2.4 Contrast-invariant and foundation models

SynthMorph [37] achieves contrast invariance by training on images synthesized from label maps, so the network never sees a real acquired image during training. BrainMorph [38] is a keypoint-based foundation model for brain MRI, and GradICON, uniGradICON, and MultiGradICON [39, 9, 10] pursue task-agnostic generality through gradient inverse-consistency regularization and large-scale pretraining, and can be applied zero-shot. Our approach differs from these in two ways. First, we compare a learned descriptor with a strong hand-crafted alternative under a controlled, resolution-matched protocol. Second, we show that combining the two is better than either, a result that, to our knowledge, has not previously been quantified for multimodal brain MRI.

3 Methods

The method has two stages (Fig. 1). In Stage 1, CMDN is trained by self-supervision to produce contrast-invariant voxel embeddings and is then frozen. In Stage 2, a diffeomorphic registration network is trained with a hybrid similarity that combines the frozen CMDN embedding with MIND-SSC. Only the registration network is used at test time.

Unsupervised Diffeomorphic Multimodal MRI Registration with a Learned Cross-Modal Descriptor

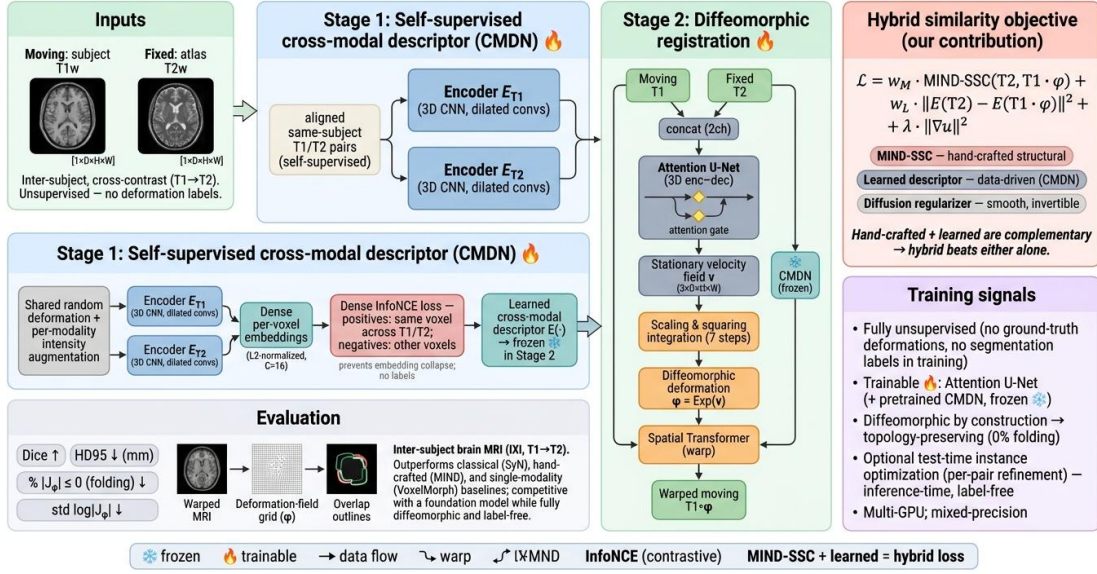


Figure 1. Overview of the proposed framework. *Stage 1*: CMDN learns a contrast-invariant descriptor by dense contrastive self-supervision and is then frozen. *Stage 2*: a diffeomorphic attention U-Net is trained with the hybrid (learned + MIND-SSC) similarity and a diffusion regularizer. No anatomical labels are used, and inference is a single forward pass of the U-Net.

3.1 Problem formulation and notation

Let $I_m, I_f: \Omega \subset \mathbb{Z}^3 \rightarrow \mathbb{R}$ denote the moving (T1w) and fixed (T2w) volumes defined on a discrete grid Ω . We seek a spatial transform $\phi: \Omega \rightarrow \mathbb{R}^3$ such that the warped image $I_m \circ \phi$ is anatomically aligned with I_f . We write $\phi = \text{Id} + u$, where u is the displacement field.

Because I_m and I_f have different contrasts, alignment cannot be measured by comparing intensities directly; instead we compare their representations under contrast-invariant descriptors, as described below.

3.2 Diffeomorphic registration backbone

A compact 3D attention U-Net [40, 41], g_θ , takes the concatenated pair (I_m, I_f) and predicts a stationary velocity field $v = g_\theta(I_m, I_f)$. The deformation is the group exponential of this field, $\phi = \exp(v)$, obtained by scaling and squaring [22]:

$$\phi^{(1/2^K)} = \text{Id} + \frac{v}{2^K}, \quad \phi^{(1/2^{k-1})} = \phi^{(1/2^k)} \circ \phi^{(1/2^k)}, \quad k = K, \dots, 1, \quad (1)$$

with $\phi = \phi^{(1)}$ and $K = 7$. For a sufficiently smooth velocity field, the resulting map is diffeomorphic: it is smooth, invertible, and topology preserving, and its inverse is obtained by integrating $-v$. Warping uses a differentiable spatial transformer [42] with trilinear interpolation, so the whole pipeline can be trained end to end.

Attention gates [41] on the skip connections of g_θ let the decoder focus on structurally informative regions. The network has 0.247 M trainable parameters in total. No ground-truth deformations or segmentations are used in training, and no descriptor needs to be computed at test time.

3.3 Cross-Modal Descriptor Network (CMDN)

CMDN consists of two modality-specific encoders, E_{T1} and E_{T2} . Each is a dilated 3D convolutional network that maps a volume to an L_2 -normalized C -dimensional embedding at every voxel ($C = 16$ by default). Dilated convolutions give a large receptive field while preserving full spatial resolution, which is needed for a dense descriptor. The two encoders have separate weights because the two contrasts require different low-level filters to reach a shared representation.

Generating positive pairs

CMDN is trained on aligned same-subject T1w/T2w pairs. For such a pair, the voxel at location p in the T1w image and the voxel at p in the T2w image show the same tissue. We create two training views by applying one *shared* random smooth deformation to both contrasts, which preserves voxelwise correspondence while exposing the encoders to varied anatomy and geometry. Each contrast then receives an independent intensity perturbation (random gamma, a smooth multiplicative bias field, and additive Gaussian noise), so that the encoders cannot rely on scanner-specific intensity statistics.

Dense contrastive objective

Let $\tilde{I}_{T1}$ and $\tilde{I}_{T2}$ be the augmented views and let $z^a = E_{T1}(\tilde{I}_{T1})$ and $z^b = E_{T2}(\tilde{I}_{T2})$. We sample S voxel locations inside the brain mask and use a symmetric dense InfoNCE loss [35, 36], which pulls together embeddings at matching locations and pushes apart embeddings at all other sampled locations:

$$\mathcal{L}_{\text{NCE}} = -\frac{1}{2S} \sum_{i=1}^S \left[\log \frac{e^{\langle z_i^a, z_i^b \rangle / \tau}}{\sum_{j=1}^S e^{\langle z_i^a, z_j^b \rangle / \tau}} + \log \frac{e^{\langle z_i^b, z_i^a \rangle / \tau}}{\sum_{j=1}^S e^{\langle z_i^b, z_j^a \rangle / \tau}} \right], \quad (2)$$

where $\langle \cdot, \cdot \rangle$ is the inner product, which equals cosine similarity for unit-norm embeddings, and τ is a temperature. The negatives give the objective a non-trivial optimum. Without them, the trivial solution of mapping every voxel to the same vector would minimize a pure alignment loss, a collapse that makes naive learned similarities degenerate. After pretraining, CMDN is frozen and used only as a fixed similarity in Stage 2.

3.4 MIND-SSC descriptor

For completeness we summarize the hand-crafted descriptor [3, 19]. For a voxel x and a set $\mathcal{R}$ of neighbour offsets, MIND is

$$\text{MIND}(I, x, r) = \frac{1}{Z} \exp \left(-\frac{D_p(I, x, x + r)}{V(I, x)} \right), \quad r \in \mathcal{R}, \quad (3)$$

where D_p is the sum of squared differences between patches centered at its two arguments, $V(I, x)$ is a local variance estimate, and Z normalizes the descriptor so that its maximum is one. The self-similarity-context variant (MIND-SSC) computes D_p between pairs of patches in the six-neighbourhood of x rather than between x and its neighbours, which makes it more

robust to noise and to the central voxel. The MIND-SSC dissimilarity between two images is the mean absolute difference of their descriptor vectors over all voxels and channels.

3.5 Hybrid similarity and training objective

Our central hypothesis is that the learned CMDN embedding and the hand-crafted MIND-SSC descriptor capture *complementary* cross-modal structure. We therefore train the registration network to minimize

$$\begin{aligned} \mathcal{L}(\theta) = & w_M \mathcal{D}_{\text{MIND}}(I_f, I_m \circ \phi) \\ & + w_L \frac{1}{|\Omega|} \sum_{x \in \Omega} \|E_{T_2}(I_f)(x) - E_{T_1}(I_m \circ \phi)(x)\|_2^2 \\ & + \lambda \frac{1}{|\Omega|} \sum_{x \in \Omega} \|\nabla u(x)\|^2, \end{aligned} \tag{4}$$

where $\mathcal{D}_{\text{MIND}}$ is the MIND-SSC dissimilarity, the second term compares CMDN embeddings of the fixed image and of the warped moving image, and the third term is a diffusion regularizer on the displacement $u = \phi - \text{Id}$ that encourages smooth deformations. Every term is label-free. Setting $w_L = 0$ gives the MIND-SSC baseline in the same backbone, and setting $w_M = 0$ gives the learned-descriptor variant, so that both ablations differ from the full model in the similarity term only.

3.6 Training and inference

Training proceeds in two stages. Stage 1 trains CMDN with Eq. (2). Stage 2 freezes CMDN and trains g_θ with Eq. (4). The registration network is trained with the Adam optimizer (learning rate 10^{-4}) and mixed-precision arithmetic. At inference, a moving and a fixed volume are passed once through g_θ , the velocity field is integrated with Eq. (1), and the moving image (or its label map) is warped. Neither CMDN nor MIND-SSC is evaluated at test time.

4 Experimental setup

4.1 Datasets and preprocessing

IXI

The IXI dataset [43] provides co-acquired T1w and T2w brain MRI from healthy adults scanned at three London hospitals with different scanners: Guy’s Hospital (Guys), Hammersmith Hospital (HH), and the Institute of Psychiatry (IOP). We evaluate *inter-subject* registration to a fixed atlas: the T1w image of each test subject is aligned to a held-out atlas T2w image. This follows the standard protocol for deformable brain registration [5]. We chose it because intra-subject, same-session pairs are already nearly aligned after an affine step and therefore do not discriminate between deformable methods. The test set contains 45 subjects, and no test subject is used for training.

Preprocessing

All volumes are resampled to 1 mm isotropic resolution, corrected for intensity inhomogeneity with N4, affinely aligned to the atlas, and cropped to $144 \times 160 \times 192$ voxels.

Scanner domain shift

To test generalization across scanners, we train on two sites (Guys and HH) and test on all available subjects from the unseen IOP site ($n = 73$).

Pathology

To test generalization to abnormal anatomy, we use T1 and T2 volumes from glioma patients in the BraTS dataset [44] ($n = 34$). These cases combine large anatomical abnormality, abnormal tissue appearance, and tumor mass effect, none of which is present in the IXI training data. Preprocessing follows the IXI protocol.

Evaluation labels

Anatomical labels for evaluation are produced by the contrast-agnostic segmentation tool SynthSeg [45], applied separately to both contrasts. These labels are used *only* for scoring and never during training.

4.2 Baselines

We compare against representatives of each family discussed in Section 2:

- **SyN** [16] (ANTs), a classical iterative diffeomorphic method with a mutual-information similarity.
- **VoxelMorph** [4], the standard unsupervised learning method, trained with its native local NCC loss.
- **MIND-SSC** [3, 19], the hand-crafted descriptor used as the training similarity within our backbone ($w_L = 0$ in Eq. (4)).
- **Learned (ours)**, the CMDN descriptor alone within our backbone ($w_M = 0$).
- **NestedMorph** [26], a recent nested-attention transformer, trained with a MIND objective at its native 128^3 input size.
- **CorrMorph** [30], a recent 118 M-parameter correlation-mining CNN–transformer, trained with its native NCC objective.
- **MultiGradICON** [10], a multimodal registration foundation model, applied zero-shot with the released weights.

Unless stated otherwise, trainable baselines use the same data splits, preprocessing, and training budget as our method.

4.3 Evaluation metrics

Registration accuracy is measured by the Dice similarity coefficient and the 95th percentile Hausdorff distance (HD95) between the warped moving labels and the fixed labels, averaged over SynthSeg structures. Deformation regularity is measured by the percentage of voxels

with non-positive Jacobian determinant, $\%|J_\phi| \leq 0$, where $J_\phi(x) = \nabla\phi(x)$, which indicates folding, and by the standard deviation of $\log|J_\phi|$, which reflects how strongly the deformation compresses and expands tissue.

4.4 Statistical analysis and evaluation protocol

Differences between methods are tested with the two-sided paired Wilcoxon signed-rank test on per-subject scores, with significance at $p < 0.05$. We also report, for each baseline, the percentage of test subjects on which the hybrid attains a higher Dice. *All methods are scored at an identical 144^3 field of view.* During development we found that differences in evaluation resolution are a subtle but material confounder, able to shift Dice by amounts comparable to the differences between methods, so we control for it strictly.

4.5 Implementation

All models are trained on NVIDIA RTX 2080 Ti GPUs (11 GB). Code and trained models will be released publicly on acceptance.

4.6 Ethics approval and consent

Ethics approval and consent to participate were not applicable to this secondary analysis of publicly available IXI and BraTS MRI datasets.

5 Results

5.1 Comparison with existing methods

Table 1 reports the resolution-matched comparison on IXI. The hybrid achieves the highest Dice (**0.700**), the lowest HD95 (**3.21** mm), and no folded voxels, and it significantly outperforms every baseline tested ($p < 0.05$, paired Wilcoxon).

The results fall into clear tiers according to the similarity used for training. At the bottom, the single-modality NCC objective (VoxelMorph) essentially fails across contrasts (Dice 0.372): local cross-correlation has no consistent optimum when white matter and CSF swap their relative brightness between T1w and T2w. The classical mutual-information optimizer SyN is diffeomorphic but limited in accuracy (0.470) and slow. Structural and learned objectives form the upper tier: MIND-SSC reaches 0.599, the CMDN descriptor 0.664, and their combination 0.700. The foundation model MultiGradICON, applied zero-shot, reaches 0.591. It is competitive on large structures but, as Section 5.2 shows, performs poorly on small deep nuclei.

Among learning backbones, the recent transformers underperform our compact network under the matched protocol. NestedMorph, trained with a MIND objective at its native 128^3 input size, reaches 0.525. The 118 M-parameter CorrMorph, trained with its native cross-correlation objective, reaches 0.658, second only to our descriptor variants. It does so, however, with a transform that is far from diffeomorphic: on average 8.9% of voxels fold, and $\text{stdlog}|J| = 0.99$, compared with no folding for the hybrid. CorrMorph’s backbone is roughly 470 times larger than ours, and with a monomodal objective it gains overlap by folding the deformation rather than by finding valid correspondences. The hybrid is therefore best on both accuracy and deformation validity.

Table 1. Inter-subject T1→T2 registration on IXI, scored at 144^3 (mean over $n = 45$ test subjects). *: $p < 0.05$ versus Hybrid (paired Wilcoxon). Best values in bold.

Method	Dice $\uparrow$	HD95 (mm) $\downarrow$	% $ J \leq 0 \downarrow$	std log $ J $
VoxelMorph (NCC) [4]	0.372*	8.66*	0.000	0.000
SyN (ANTs) [16]	0.470*	7.45*	0.000	0.294
MultiGradICON (zero-shot) [10]	0.591*	5.67*	0.002*	0.289
MIND-SSC [3]	0.599*	4.76*	0.000	0.264
CorrMorph (NCC) [30]	0.658*	5.91*	8.90*	0.993
Learned (ours)	0.664*	3.71*	0.000	0.196
Hybrid (ours)	0.700	3.21	0.000	0.332

The advantage is also consistent across subjects. Fig. 2 shows the per-subject Dice distribution: the hybrid has the highest median (0.735) and the narrowest spread. In paired comparisons it achieves a higher Dice on 100% of test subjects against VoxelMorph, SyN, and MIND-SSC, on 82% against the learned descriptor alone ($p = 2.5 \times 10^{-8}$), and on 76% against MultiGradICON ($p = 2.4 \times 10^{-5}$). The improvement is therefore not driven by a small number of easy cases.

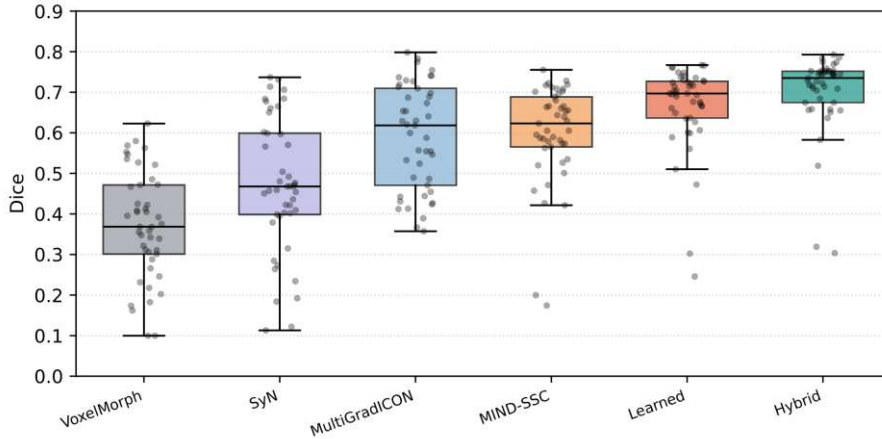


Figure 2. Per-subject Dice distribution on IXI ($n = 45$, scored at 144^3). Boxes show quartiles and dots show individual subjects. The hybrid has the highest median and the smallest dispersion.

5.2 Per-structure accuracy

Supplementary Table S1 breaks Dice down by anatomical structure, averaging left and right hemispheres over 18 SynthSeg labels. On large structures with high contrast, such as cerebral white matter and the lateral ventricles, MultiGradICON is competitive or best. On small, deep structures, however, its accuracy drops sharply (hippocampus 0.376, amygdala 0.290, pallidum 0.514), whereas the hybrid remains accurate (hippocampus 0.659, amygdala 0.658, pallidum 0.688). The hybrid achieves the best Dice on six of the nine structures and the highest mean. Its largest margins are on the deep gray-matter nuclei and the medial temporal lobe, which are the structures of most interest for morphometry in neurodegenerative disease and for targeting in functional neurosurgery. Fig. 3 shows the same data graphically.

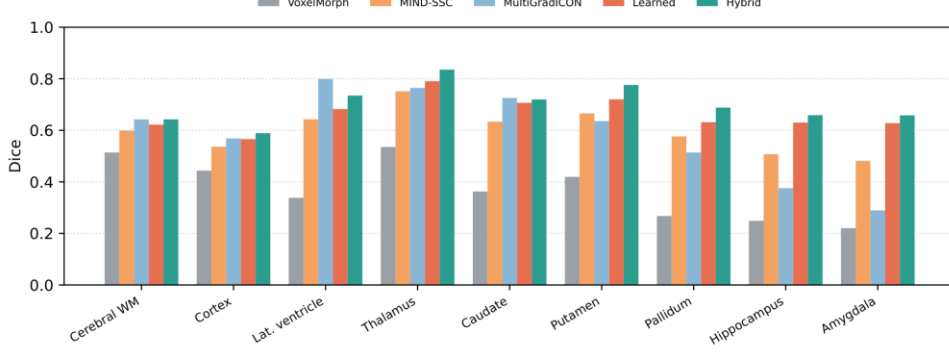


Figure 3. Per-structure Dice across methods. The hybrid’s margin is largest on deep gray-matter nuclei (thalamus, putamen, pallidum) and on the hippocampus and amygdala, where the intensity-based and foundation-model baselines degrade.

5.3 Ablation study

Descriptor complementarity

Because the MIND-SSC, learned, and hybrid variants share the same backbone, data, and training schedule and differ only in the similarity term, Table 1 also serves as an ablation of the descriptor. It shows a strict ordering: $NCC < MIND-SSC < Learned < Hybrid$. The learned descriptor improves Dice by 0.065 over MIND-SSC, and combining the two adds a further 0.036 over the learned descriptor alone. The hybrid beats the learned descriptor on 82% of test subjects ($p < 10^{-7}$). A purely learned metric can adapt too closely to the intensity statistics seen in training, while MIND-SSC encodes a fixed geometric prior; in combination, each compensates for the other’s weaknesses. The slightly higher $\text{stdlog}||$ of the hybrid (0.332 versus 0.196 for the learned descriptor) indicates that it applies larger local volume changes, which is consistent with its better recovery of small structures, while still producing no folds.

Design choices

Supplementary Table S2 examines two further design elements. Replacing the diffeomorphic (stationary velocity) parameterization with a directly predicted displacement field lowers Dice from 0.700 to 0.679 and removes the guarantee of an invertible, fold-free transform. Reducing the CMDN embedding from 16 to 8 dimensions lowers the Dice of the learned descriptor from 0.664 to 0.627, showing that a sufficiently expressive contrast-invariant embedding is needed. We keep both defaults.

5.4 Generalization to an unseen scanner and to pathology

Table 2 summarizes both generalization experiments.

Unseen scanner

When trained on Guys and HH and tested on the held-out IOP scanner ($n = 73$), the hybrid keeps its lead (Dice 0.456) over MIND-SSC (0.433), SyN (0.305), and VoxelMorph (0.262). All methods score lower than in the main experiment, as expected under a change of scanner, but the ranking is unchanged. This suggests that the learned component does not simply memorize scanner-specific intensities.

Pathology

On glioma cases from BraTS ($n = 34$), whose anatomy is distorted by tumor and mass effect and whose tissue appearance is abnormal, the hybrid again leads (Dice 0.519), followed by the learned descriptor (0.486), MIND-SSC (0.467), VoxelMorph (0.381), and SyN (0.214). It also has the lowest HD95 (5.6 mm) and no folding. The complementarity ordering (Hybrid $>$ Learned $>$ MIND-SSC) holds outside the training distribution. Fig. 4 shows a representative case. Although the tumor (outlined) compresses the ipsilateral lateral ventricle, the hybrid recovers the ventricular and midline anatomy of the atlas most faithfully, whereas VoxelMorph and, to a lesser extent, MIND-SSC leave visible residual misalignment in the checkerboard.

Table 2. Generalization to an unseen scanner (IXI IOP site, $n = 73$) and to pathology (BraTS glioma, $n = 34$). Dice $\uparrow$. Best value per setting in bold.

Setting	Method	Dice $\uparrow$
Unseen scanner (IOP)	VoxelMorph	0.262
	SyN	0.305
	MIND-SSC	0.433
	Hybrid (ours)	0.456
Pathology (BraTS)	VoxelMorph	0.381
	SyN	0.214
	MIND-SSC	0.467
	Learned (ours)	0.486
	Hybrid (ours)	0.519

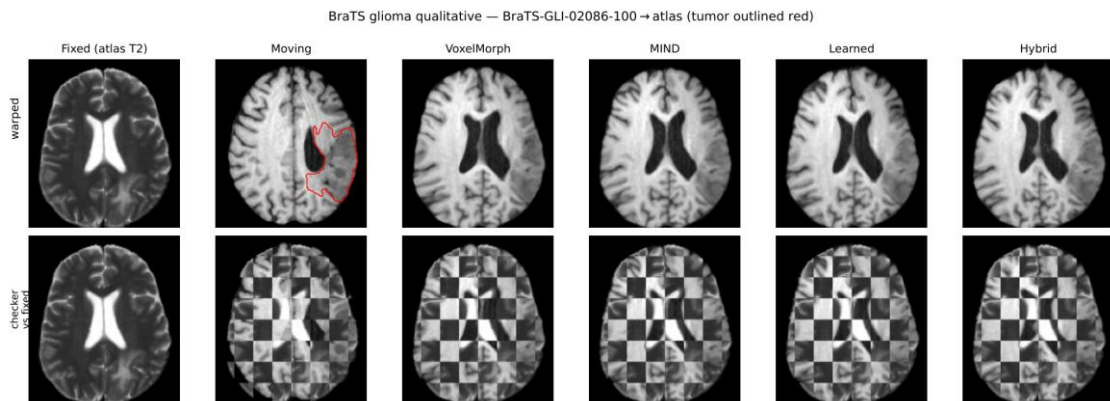


Figure 4. Qualitative results on pathology (BraTS glioma, subject $\rightarrow$ atlas). Top: warped moving image, with the tumor outlined in red on the moving image. Bottom: checkerboard against the fixed atlas (0% folding for all learning methods). Despite tumor mass effect on the lateral ventricle, the hybrid recovers ventricular and midline anatomy most faithfully; residual misalignment is visible for VoxelMorph and MIND-SSC.

5.5 Qualitative analysis

Fig. 5 shows a representative case for our method. The warped moving image conforms to the fixed atlas, the checkerboard composite is anatomically continuous across tile boundaries, the deformation grid is smooth, and the Jacobian determinant is strictly positive everywhere.

Fig. 6 compares all methods on one case, and the three rows lead to different conclusions. In the *warped-image* row (top), every method produces a plausible-looking T1w image that resembles the atlas, so warped images alone discriminate poorly between methods. The *checkerboard* row (middle) interleaves the warped moving image with the fixed atlas and exposes residual misalignment as broken anatomy at tile edges. VoxelMorph leaves clear ventricular and cortical discontinuities because it barely deforms the image. SyN and MIND-SSC sharpen boundaries but under-expand the ventricles, and MultiGradICON aligns gross anatomy while blurring deep structures. The learned and hybrid variants give the most continuous composites, especially around the lateral ventricles and deep gray matter.

The *Jacobian* row (bottom) reveals what the warped images hide. VoxelMorph’s $|J_\phi|$ is close to 1 everywhere: its NCC objective provides almost no cross-modal gradient, so it applies a near-identity transform. Its regularity is high only because it barely registers. MIND-SSC, Learned, and Hybrid apply progressively richer deformations that remain smooth and strictly positive. CorrMorph produces a warped image that looks as clean as any other, but its Jacobian is highly irregular, and 7.9% of voxels fold in this case (green contours; the mean over the test set is 8.9%, Table 1). Its plausible appearance is thus produced by a non-invertible transform. This figure illustrates the main practical point of the paper: a good-looking warped image does not imply a valid deformation, and in our experiments only the descriptor objectives with a diffeomorphic parameterization provided both.

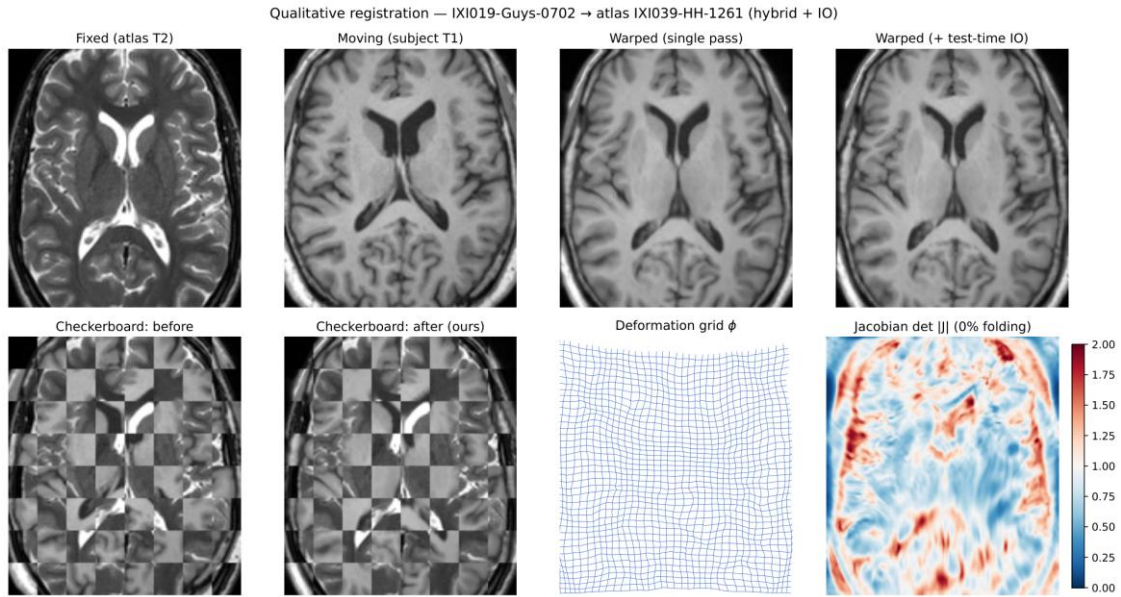


Figure 5. Qualitative registration result for the proposed method. Top: fixed, moving, warped (single forward pass), and warped with optional test-time refinement. Bottom: checkerboard before and after registration, deformation grid, and Jacobian determinant (0% folding).

other can resolve the ambiguity. This yields the best Dice and HD95, with gains concentrated on the small deep nuclei that are hardest to register. (6) The *foundation model* (MultiGradICON), pretrained on many datasets and applied zero-shot, aligns gross anatomy but is not tuned to this contrast pair and loses accuracy on the same small structures (hippocampus 0.376, amygdala 0.290). (7) The *recent transformers* show that model capacity does not substitute for an appropriate objective. NestedMorph with a MIND loss reaches 0.525, and CorrMorph with its native cross-correlation loss reaches competitive overlap (0.658) only by producing a heavily folded transform, whereas our 0.25 M network is both more accurate and strictly invertible.

Two design choices turn this better similarity signal into accuracy without compromising validity. First, the descriptors are used only during training, so they add no test-time cost. Second, the stationary velocity parameterization guarantees a topology-preserving deformation by construction, so accuracy is never bought with folding.

6.2 Clinical and practical implications

Several properties of the method matter for clinical use. First, multi-contrast analyses such as tissue segmentation, lesion quantification, and atlas-based labeling assume that each voxel refers to the same anatomy in every contrast. Errors in the deep gray-matter nuclei and medial temporal lobe are particularly consequential, because these structures are small, their volumes are used as biomarkers in neurodegenerative disease, and they are targets in stereotactic procedures. The hybrid’s largest gains are on exactly these structures.

Second, fold-free deformations are a prerequisite for propagating contours and dose distributions in radiotherapy planning and for Jacobian-based morphometry, where a negative Jacobian has no physical meaning. A method that achieves high overlap by folding, as CorrMorph does here, cannot be used safely in these settings, however good its warped images look.

Third, the training pipeline needs no manual annotations. Institutions can therefore adapt the model to their own scanners and protocols using only routinely acquired, unlabeled T1w/T2w pairs. Fourth, with a sub-second runtime and a network of fewer than 0.3 M parameters, registration can run on modest hardware within an automated processing pipeline, without dedicated high-memory GPUs.

6.3 Evaluation methodology

Two aspects of our protocol may be useful to others. First, we found that the resolution at which methods are scored can change Dice by amounts similar to the differences between methods. Comparisons in which each method is evaluated at its own native resolution can therefore be misleading, and we recommend that registration studies report and match the evaluation field of view. Second, overlap metrics should always be reported together with deformation regularity. In our experiments, a ranking by Dice alone would have placed CorrMorph close to the descriptor-based methods and hidden the fact that its transforms are not invertible.

6.4 Limitations and future work

This study has several limitations. First, CMDN builds on earlier contrastive representations for registration [6, 7, 8]. Our contribution is its combination with a hand-crafted descriptor, the evidence that the two are complementary, and the realization of this combination as a compact, label-free, diffeomorphic system. Second, the comparison with MultiGradICON is across paradigms (a zero-shot foundation model versus a network trained on the target domain); we matched the evaluation field of view and present it as a strong reference rather than a like-for-like comparison. Third, models were trained at 144^3 because of GPU memory limits. Optional test-time instance optimization gave only marginal additional gains and is not used in the reported results. Fourth, evaluation labels come from SynthSeg rather than from manual annotation. SynthSeg is accurate and contrast-agnostic, but its errors add noise to the absolute Dice values, although they affect all methods equally. Fifth, we studied a single contrast pair (T1w to T2w) in the brain. The learned-descriptor variant was not evaluated on the unseen-scanner split, and NestedMorph was evaluated at its native 128^3 input.

Future work will extend the framework to other contrast pairs, such as FLAIR and diffusion-derived maps, and to other modality pairs, such as MRI–CT, where non-monotonic intensity relationships are also common. It will also explore learning the weighting between the two descriptors adaptively in space, and evaluating the method on manually annotated data and in downstream clinical tasks such as lesion tracking and radiotherapy contour propagation.

7 Conclusion

For unsupervised multimodal brain MRI registration, the cross-modal similarity objective is the most important design choice. A frozen, label-free learned descriptor combined with the hand-crafted MIND-SSC descriptor provides complementary, contrast-invariant information. Within a compact diffeomorphic attention U-Net, this combination gave the most accurate results among the methods compared, with a guaranteed topology-preserving transform. In contrast, a transformer baseline about 470 times larger reached comparable overlap only by folding the deformation. The approach generalized across scanners and to brains with glioma, needs no labels for training, and registers a volume pair in under 0.1 s, which makes it a practical component for multi-contrast clinical pipelines.

Data availability

The MRI datasets analysed in this study are available from their original providers. The Information eXtraction from Images (IXI) Dataset is available from the Brain Development website at <https://brain-development.org/ixi-dataset/>. The BraTS 2024 BraTS-GLI training dataset, distributed as BraTS2024-BraTS-GLI-TrainingData.zip, is available from the Synapse repository under accession syn60086071, within collection syn59059776, at <https://www.synapse.org/Synapse:syn60086071>. Access to BraTS data requires a Synapse account and compliance with the applicable dataset access and data-use conditions. Access to both source datasets is subject to the respective providers’ terms and conditions.

Code availability

Code and trained models will be released publicly on acceptance.

Author contributions

D M Bappy conceptualised the study, developed the software, and wrote and edited the manuscript. Hyeongboo Baek contributed to manuscript writing, supervised the study, and arranged resources.

Funding

This research received no specific funding.

Additional information

Competing interests

The authors declare no competing interests.

Supplementary information

Supplementary Information accompanies this article and contains Supplementary Tables S1–S3.

References

- [1] Maes, F., Collignon, A., Vandermeulen, D., Marchal, G. & Suetens, P. Multimodality image registration by maximization of mutual information. *IEEE Trans. Med. Imaging* **16**, 187–198 (1997).
- [2] Wells, W. M., Viola, P., Atsumi, H., Nakajima, S. & Kikinis, R. Multi-modal volume registration by maximization of mutual information. *Med. Image Anal.* **1**, 35–51 (1996).
- [3] Heinrich, M. P. *et al.* MIND: modality independent neighbourhood descriptor for multi-modal deformable registration. *Med. Image Anal.* **16**, 1423–1435 (2012).
- [4] Balakrishnan, G., Zhao, A., Sabuncu, M. R., Guttag, J. & Dalca, A. V. VoxelMorph: a learning framework for deformable medical image registration. *IEEE Trans. Med. Imaging* **38**, 1788–1800 (2019).
- [5] Chen, J. *et al.* TransMorph: transformer for unsupervised medical image registration. *Med. Image Anal.* **82**, 102615 (2022).
- [6] Pielawski, N. *et al.* CoMIR: contrastive multimodal image representation for registration. In *Advances in Neural Information Processing Systems (NeurIPS)* (2020).
- [7] Dey, N. *et al.* ContraReg: contrastive learning of multi-modality unsupervised deformable image registration. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, 66–77 (2022).

- [8] Mok, T. C. W. *et al.* Modality-agnostic structural image representation learning for deformable multi-modality medical image registration. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 11215–11225 (2024).
- [9] Tian, L. *et al.* uniGradICON: a foundation model for medical image registration. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)* (2024).
- [10] Demir, B. *et al.* MultiGradICON: a foundation model for multimodal medical image registration. In *Biomedical Image Registration (WBIR)*, 3–18 (2024).
- [11] Sotiras, A., Davatzikos, C. & Paragios, N. Deformable medical image registration: a survey. *IEEE Trans. Med. Imaging* **32**, 1153–1190 (2013).
- [12] Haskins, G., Kruger, U. & Yan, P. Deep learning in medical image registration: a survey. *Mach. Vis. Appl.* **31**, 8 (2020).
- [13] Rueckert, D. *et al.* Nonrigid registration using free-form deformations: application to breast MR images. *IEEE Trans. Med. Imaging* **18**, 712–721 (1999).
- [14] Beg, M. F., Miller, M. I., Trounev, A. & Younes, L. Computing large deformation metric mappings via geodesic flows of diffeomorphisms. *Int. J. Comput. Vis.* **61**, 139–157 (2005).
- [15] Vercauteren, T., Pennec, X., Perchant, A. & Ayache, N. Diffeomorphic demons: efficient non-parametric image registration. *NeuroImage* **45**, S61–S72 (2009).
- [16] Avants, B. B., Epstein, C. L., Grossman, M. & Gee, J. C. Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain. *Med. Image Anal.* **12**, 26–41 (2008).
- [17] Modat, M. *et al.* Fast free-form deformation using graphics processing units. *Comput. Methods Programs Biomed.* **98**, 278–284 (2010).
- [18] Rivaz, H., Karimaghloo, Z. & Collins, D. L. Self-similarity weighted mutual information: a new nonrigid image registration metric. *Med. Image Anal.* **18**, 343–358 (2014).
- [19] Heinrich, M. P., Jenkinson, M., Papież, B. W., Brady, S. M. & Schnabel, J. A. Towards realtime multimodal fusion for image-guided interventions using self-similarities. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, 187–194 (2013).
- [20] Balakrishnan, G., Zhao, A., Sabuncu, M. R., Guttag, J. & Dalca, A. V. An unsupervised learning model for deformable medical image registration. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 9252–9260 (2018).
- [21] de Vos, B. D. *et al.* A deep learning framework for unsupervised affine and deformable image registration. *Med. Image Anal.* **52**, 128–143 (2019).
- [22] Dalca, A. V., Balakrishnan, G., Guttag, J. & Sabuncu, M. R. Unsupervised learning of probabilistic diffeomorphic registration for images and surfaces. *Med. Image Anal.* **57**, 226–236 (2019).
- [23] Mok, T. C. W. & Chung, A. C. S. Large deformation diffeomorphic image registration with Laplacian pyramid networks. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, 211–221 (2020).

- [24] Kim, B. *et al.* CycleMorph: cycle consistent unsupervised deformable image registration. *Med. Image Anal.* **71**, 102036 (2021).
- [25] Wang, H., Ni, D. & Wang, Y. ModeT: learning deformable image registration via motion decomposition transformer. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)* (2023).
- [26] Kumar, G. M. K., Mendola, J. & Shmuel, A. NestedMorph: enhancing deformable medical image registration with nested attention mechanisms. In *Proceedings of the Winter Conference on Applications of Computer Vision (WACV)*, 4683–4692 (2025).
- [27] Jia, X. *et al.* Fourier-Net: fast image registration with band-limited deformation. In *AAAI Conference on Artificial Intelligence* (2023).
- [28] Meng, M., Feng, D., Bi, L. & Kim, J. Correlation-aware coarse-to-fine MLPs for deformable medical image registration. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2024).
- [29] Wang, H., Ni, D. & Wang, Y. Recursive deformable pyramid network for unsupervised medical image registration. *IEEE Trans. Med. Imaging* **43**, 2229–2240 (2024).
- [30] Chang, Y., Li, Z. & Yang, N. CorrMorph: unsupervised deformable brain MRI registration based on correlation mining. *IEEE J. Biomed. Health Inform.* **29**, 2798–2807 (2025).
- [31] Hering, A. *et al.* Learn2Reg: comprehensive multi-task medical image registration challenge, dataset and evaluation in the era of deep learning. *IEEE Trans. Med. Imaging* **42**, 697–712 (2023).
- [32] Arar, M., Ginger, Y., Danon, D., Bermanno, A. H. & Cohen-Or, D. Unsupervised multi-modal image registration via geometry preserving image-to-image translation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2020).
- [33] Fan, J., Cao, X., Wang, Q., Yap, P. T. & Shen, D. Adversarial learning for mono- or multi-modal registration. *Med. Image Anal.* **58**, 101545 (2019).
- [34] Qin, C. *et al.* Unsupervised deformable registration for multi-modal images via disentangled representations. In *Information Processing in Medical Imaging (IPMI)* (2019).
- [35] van den Oord, A., Li, Y. & Vinyals, O. Representation learning with contrastive predictive coding. Preprint at <https://arxiv.org/abs/1807.03748> (2018).
- [36] Chen, T., Kornblith, S., Norouzi, M. & Hinton, G. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning (ICML)* (2020).
- [37] Hoffmann, M. *et al.* SynthMorph: learning contrast-invariant registration without acquired images. *IEEE Trans. Med. Imaging* **41**, 543–558 (2022).
- [38] Wang, A. Q. *et al.* BrainMorph: a foundational keypoint model for robust and flexible brain MRI registration. Preprint at <https://arxiv.org/abs/2405.14019> (2024).

- [39] Tian, L. *et al.* GradICON: approximate diffeomorphisms via gradient inverse consistency. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2023).
- [40] Ronneberger, O., Fischer, P. & Brox, T. U-Net: convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, 234–241 (2015).
- [41] Oktay, O. *et al.* Attention U-Net: learning where to look for the pancreas. Preprint at <https://arxiv.org/abs/1804.03999> (2018).
- [42] Jaderberg, M., Simonyan, K., Zisserman, A. & Kavukcuoglu, K. Spatial transformer networks. In *Advances in Neural Information Processing Systems (NeurIPS)* (2015).
- [43] IXI Dataset. Information eXtraction from Images. <https://brain-development.org/ixi-dataset/>.
- [44] Baid, U. *et al.* The RSNA-ASNR-MICCAI BraTS 2021 benchmark on brain tumor segmentation and radiogenomic classification. Preprint at <https://arxiv.org/abs/2107.02314> (2021).
- [45] Billot, B. *et al.* SynthSeg: segmentation of brain MRI scans of any contrast and resolution without retraining. *Med. Image Anal.* **86**, 102789 (2023).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [ScientificReportsSupplementaryInformation.docx](#)