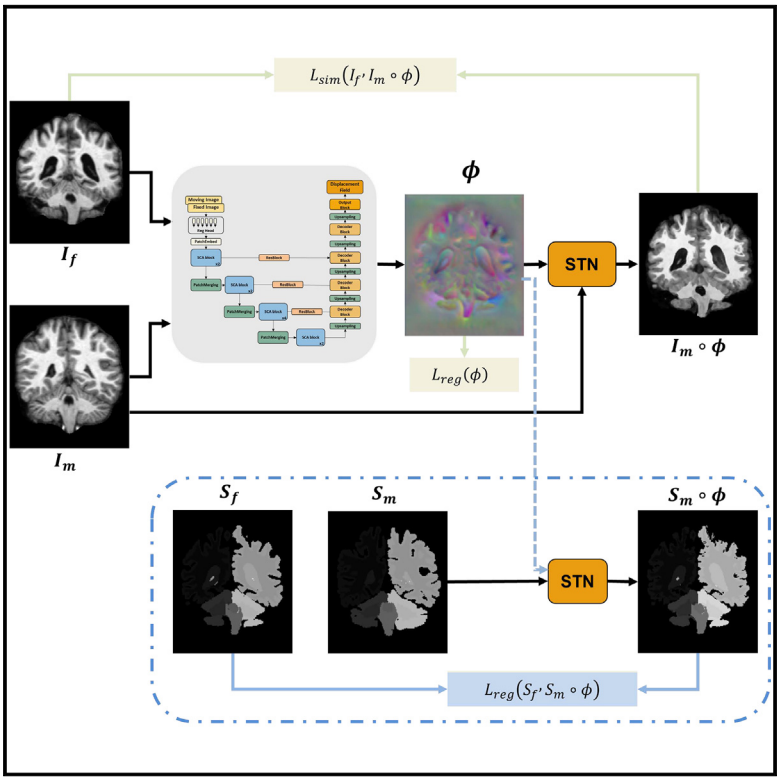
iScience

Article

Enhancing unsupervised learning in medical image registration through scale-aware context aggregation

Graphical abstract



Highlights

- ScaMorph fuses CNN and transformer for deformable image registration, enhancing accuracy
- Introduces scale-aware context aggregation for precise multiscale context assimilation
- Validated on brain MRI and CT scenarios, outperforming existing registration methods

Authors

Yuchen Liu, Ling Wang, Xiaolin Ning, Yang Gao, Defeng Wang

Correspondence

yanggao@buaa.edu.cn (Y.G.), dfwang@buaa.edu.cn (D.W.)

In brief

Medical imaging; Clinical neuroscience; Bioinformatics

Liu et al., 2025, iScience *28*, 111734
February 21, 2025
https://doi.org/10.1016/j.isci.2024.111734

CellPress

iScience

CellPress
OPEN ACCESS

Article

Enhancing unsupervised learning in medical image registration through scale-aware context aggregation

Yuchen Liu,[1] Ling Wang,[2] Xiaolin Ning,[1,2,3] Yang Gao,[1,2,3,4,*] and Defeng Wang[1,*]
[1]School of Instrumentation Science and Opto-electronics Engineering, Beihang University, Beijing 100191, China
[2]Institute of Large-Scale Scientific Facility and Centre for Zero Magnetic Field Science, Beihang University, Beijing 100191, China
[3]Hefei National Laboratory, Hefei 230000, China
[4]Lead contact
*Correspondence: yanggao@buaa.edu.cn (Y.G.), dfwang@buaa.edu.cn (D.W.)
https://doi.org/10.1016/j.isci.2024.111734

SUMMARY

Deformable image registration (DIR) is essential for medical image analysis, facilitating the establishment of dense correspondences between images to analyze complex deformations. Traditional registration algorithms often require significant computational resources due to iterative optimization, while deep learning approaches face challenges in managing diverse deformation complexities and task requirements. We introduce ScaMorph, an unsupervised learning model for DIR that employs scale-aware context aggregation, integrating multiscale mixed convolution with lightweight multiscale context fusion. This model effectively combines convolutional networks and vision transformers, addressing various registration tasks. We also present diffeomorphic variants of ScaMorph to maintain topological deformations. Extensive experiments on 3D medical images across five applications—atlas-to-patient and inter-patient brain magnetic resonance imaging (MRI) registration, inter-modal brain MRI registration, inter-patient liver computed tomography (CT) registration as well as inter-modal abdomen MRI-CT registration—demonstrate that our model significantly outperforms existing methods, highlighting its effectiveness and broader implications for enhancing medical image registration techniques.

INTRODUCTION

Deformable image registration (DIR) plays a vital role in medical imaging, facilitating the quantification of anatomical changes in longitudinal studies,[1] the fusion of multi-modal images,[2] and the alignment of patient images for treatment planning.[3] By aligning imaging data from multiple patients, physicians gain profound insights into disease characteristics and changes, ultimately enhancing patient diagnosis and treatment.[4] While traditional registration methods within a variational framework have found widespread application in medical imaging research,[5–8] they are not without limitations. Computational costs, most notably for large-scale image datasets, are of concern. Furthermore, traditional methods may struggle to accommodate significant appearance differences in image pairs due to the complexity and diversity of medical imaging data. Manual parameters and constraint selection in traditional approaches can also lead to unstable results.

In recent years, deep learning methods, particularly convolutional neural networks (CNNs),[9–12] have demonstrated significant advancements in medical image registration, leveraging deep neural networks to extract high-level features from data and enable swift and accurate image registration. Despite offering faster runtime and flexible feature representation capabilities compared to conventional optimization-based approaches, CNNs are limited in capturing long-range spatial relationships. Transformer networks,[13,14] derived from natural language processing, have shown promise in addressing this limitation and have successed in computer vision tasks. However, their application in medical image registration tasks involving significant displacements and high computational complexity remains challenging.

Motivated by the concept of convolutional modulation[15–18] and its adaptive aggregation of contextual information, we propose an innovative convolutional modulation approach named scale-aware context aggregation (SCA), which comprises two modules: multi-scale mixed convolution (MMC) and multi-scale context fusion (MCF). These modules aim to address the limitations of traditional methods by enhancing feature extraction and integration across different scales in medical image registration. To overcome the challenge of capturing long-range dependencies, a hybrid modulation transformer architecture is introduced, leveraging SCA as the encoder and a ConvNet decoder to generate a dense displacement field. This architecture effectively captures spatial correspondences between moving and fixed images, enhancing the performance of downstream tasks. Referred to as scale-aware context aggregation morph (ScaMorph), our proposed architecture encompasses a diffeomorphic variant (ScaMorph-diff) to ensure smooth and

Table 1. Quantitative results of different registration methods on the atlas-to-patient brain MR registration task

Model	DICE ↑	CI	$\lvert J_\phi\rvert \leq 0$ ↓
Affine	0.412 ± 0.207	[0.318 0.510]***	–
SyN	0.645 ± 0.152	[0.579 0.703]***	0.001 ± 0.002
NiftyReg	0.640 ± 0.166	[0.569 0.700]***	0.020 ± 0.046
LDDMM	0.680 ± 0.135	[0.619 0.732]***	0.000 ± 0.000
deedsBCV	0.733 ± 0.126	[0.671 0.781]***	0.147 ± 0.050
VoxelMorph	0.730 ± 0.127	[0.670 0.776]***	1.593 ± 0.337
TransMorph	0.752 ± 0.126	[0.690 0.800]**	1.515 ± 0.335
LKU-Net	0.751 ± 0.133	[0.685 0.799]***	3.263 ± 1.188
GroupMorph	0.738 ± 0.133	[0.673 0.787]***	2.462 ± 0.748
ScaMorph	**0.755 ± 0.132**	[**0.690 0.803**]	**1.400 ± 0.329**
VoxelMorph-diff	0.751 ± 0.128	[0.697 0.800]***	0.008 ± 0.006
TransMorph-diff	0.757 ± 0.127	[0.694 0.802]*	0.016 ± 0.009
LKU-Net-diff	0.755 ± 0.131	[0.687 0.803]**	**0.000 ± 0.000**
GroupMorph-diff	0.743 ± 0.128	[0.680 0.791]***	**0.000 ± 0.000**
RDP	0.749 ± 0.128	[0.686 0.798]***	0.013 ± 0.005
ScaMorph-diff	**0.758 ± 0.129**	[**0.700 0.808**]	0.019 ± 0.011

Dice score and percentage of voxels with a non-positive Jacobian determinant are evaluated for different methods. A 95% confidence interval (CI) was calculated for each method, with Cohen's d values indicated as follows: ∗∗∗: d ≱ 0.8, ∗∗: d ≱ 0.5, ∗: d ≱ 0.2. The bolded numbers indicate the highest scores. Data are represented as mean ± SEM.

topology-preserving deformations. Extensive qualitative and quantitative evaluations demonstrate the robustness and effectiveness of our proposed method in various brain MRI registration scenarios, including atlas-to-patient, inter-patient, and inter-modal registration, inter-patient CT registration as well as inter-modal abdomen MRI-CT registration. Comparison with existing registration methods confirms the state-of-the-art performance of our model, while ablation studies further validate the effectiveness of our proposed enhancements.

In summary, the contributions of this paper are as follows.

(1) We present ScaMorph, an innovative fusion of convolutional neural network and transformer architecture designed for deformable image registration. This pioneering model seamlessly integrates local and global dependencies, improving accuracy and resilience during registration
(2) To augment the performance of our model, we introduce the scale-aware context aggregation (SCA) mechanism, combining the powerful multiscale mixed convolution (MMC) with the evolved lightweight multiscale context fusion (MCF) to assimilate multiscale contexts, leading to more precise and reliable registration outcomes
(3) To validate the efficacy of our proposed registration model, we assess its performance across three distinct brain MRI registration scenarios, one inter-patient CT registration scenario and inter-modal MR-CT registration scenario, thoroughly compared against existing registration methods, showcasing unparalleled state-of-the-art performance. Conducting ablation studies to substantiate and affirm the effectiveness of our proposed enhancements

RELATED WORK

Deformable image registration

Deformable image registration plays a pivotal role in medical image analysis, aiming to establish voxel-level correspondence between a source image and a target image in a non-linear fashion. Existing registration algorithms utilize energy function optimization to refine the transformation progressively. Symbolically, the fixed and moving images are denoted as $\boldsymbol{I}_f$ and $\boldsymbol{I}_m$ respectively, with the registration field ϕ mapping the coordinates of $\boldsymbol{I}_f$ to those of $\boldsymbol{I}_m$. The optimization problem can be formulated as follows:

$$\widehat{\phi} = \arg\min_{\phi} \mathcal{L}(\boldsymbol{I}_f, \boldsymbol{I}_m, \phi) = \arg\min_{\phi} \mathcal{L}_{\text{sim}}(\boldsymbol{I}_f, \boldsymbol{I}_m \circ \phi) + \lambda \mathcal{L}_{\text{reg}}(\phi), \quad \text{(Equation 1)}$$

where $\boldsymbol{I}_m \circ \phi$ represents the warping of $\boldsymbol{I}_m$ by ϕ, and $\mathcal{L}_{sim}(\ldots,\ldots)$ quantifies the similarity between the two input images, commonly measured using mean squared error (MSE),[6,19] normalized cross-correlation (NCC),[7] structural similarity index (SSIM),[20,21] or mutual information (MI).[22] These metrics provide an evaluation of the likeness between the source and target images. The term $\mathcal{L}_{\text{reg}}(\phi)$ enforces regularization to foster spatial smoothness in the deformation field.

Conventional image registration methods

In the realm of medical image registration, conventional methods have primarily focused on optimizing transformations through iterative minimization of the energy function, specifically Equation 1. These methods can be broadly categorized into feature-based and anatomy-based, which measure the similarity between source and target images differently. Feature-based methods employ volume-based techniques to calculate the correlation of intrinsic features across different images, with mutual information being a commonly used approach that iteratively optimizes the transformation matrix.[7,23,24] On the other hand, anatomy-based methods leverage anatomical information and image landmarks, often implementing Random Forest, or other machine learning techniques.[6,25,26] Additionally, extensive research has been conducted on deformable registration methods based on elastic models.[8] However, the computational complexity of these traditional methods remains a hurdle due to the high cost of the iterative optimization process. Despite efforts to accelerate registration using GPU parallel computation, limitations still exist in efficiently processing large-scale medical images.

Convolutions-based image registration methods

Recent years have witnessed a surge in the popularity of deep learning-based image registration methods, particularly those employing CNNs.[27–29] These approaches harness the power of deep neural networks to extract high-level features from data, enabling efficient and precise image registration. Unlike traditional optimization-based methods, deep learning techniques offer expedited runtime and enhanced flexibility in feature

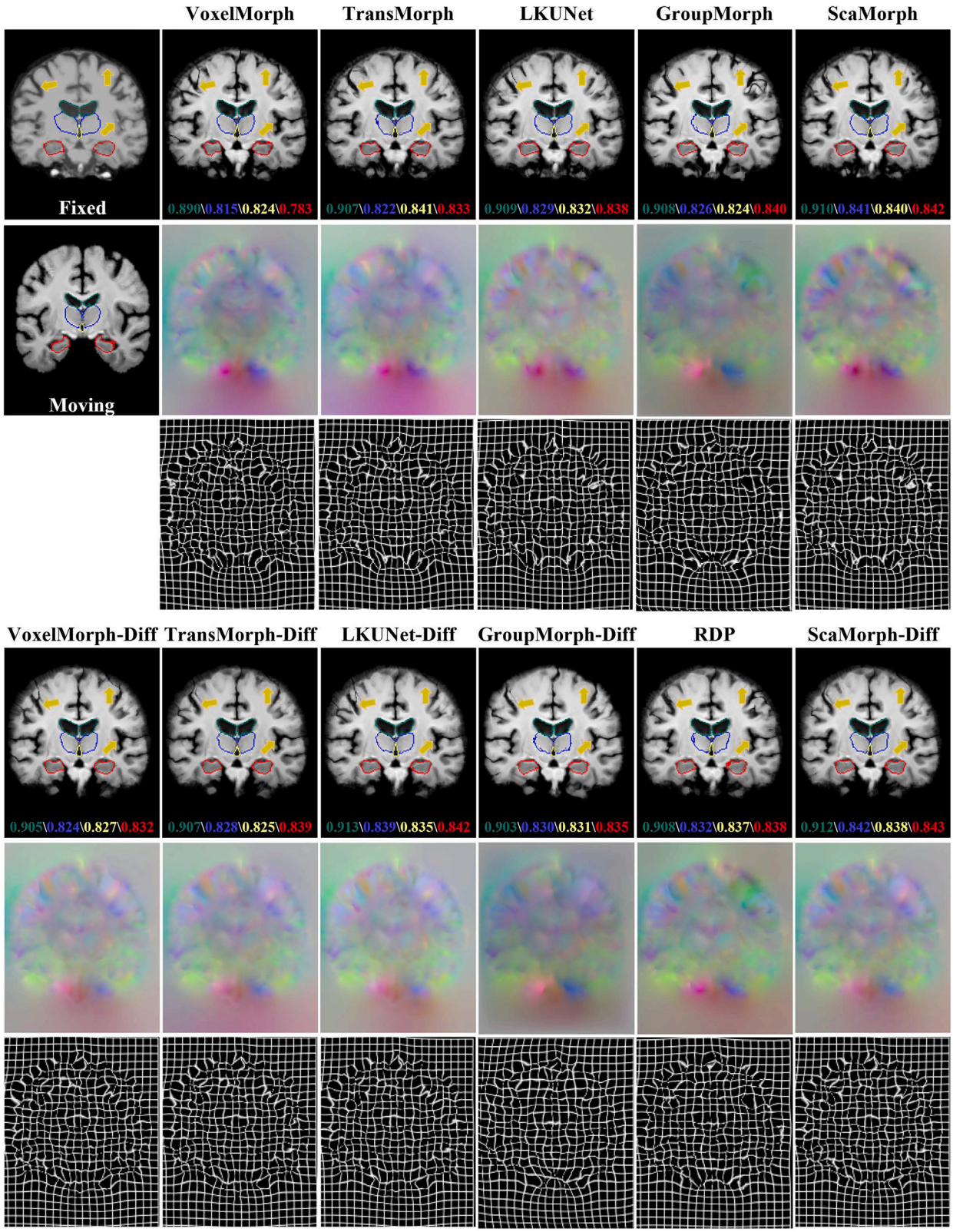


Figure 1. Qualitative results of different registration methods on the atlas-to-patient brain MR registration task

The first row displays the deformed moving images, the second row visualizes the deformation fields, and the last row illustrates the deformed grids. Specifically, the green, yellow, blue, and red contours respectively represent the ventricles, third ventricle, thalami, and hippocampi.

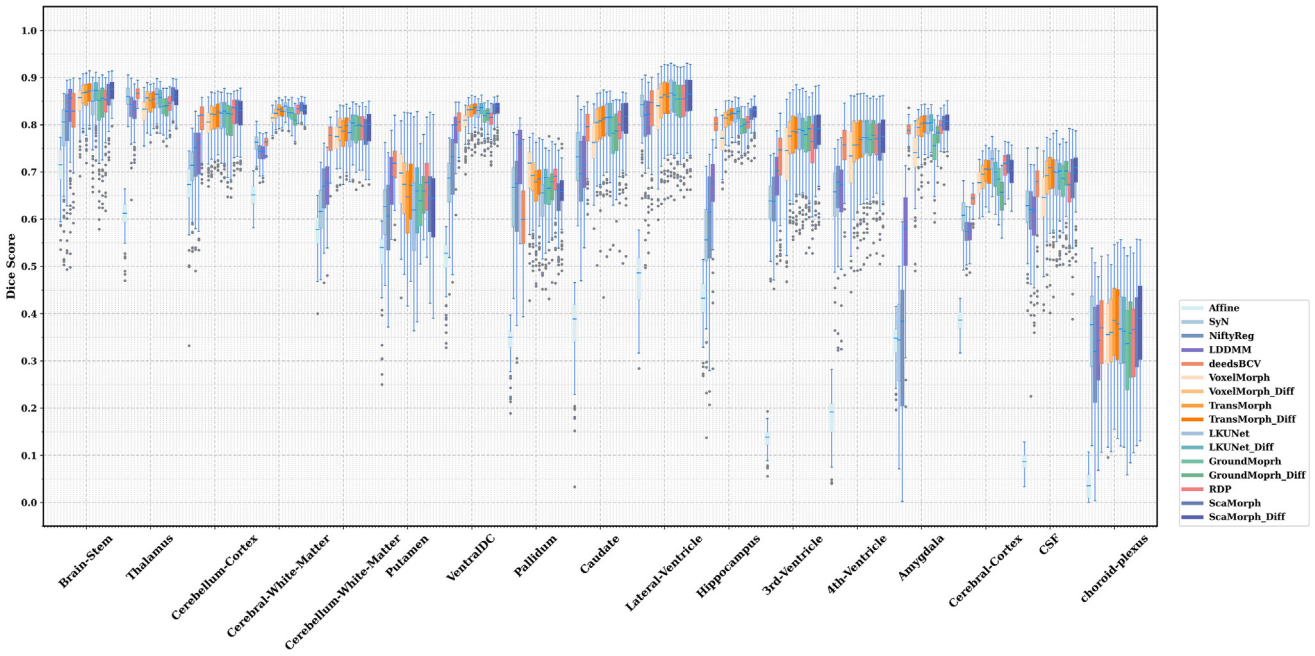


Figure 2. Quantitative comparison of different registration methods on the atlas-to-patient brain MR registration task

Boxplots depict Dice scores for different brain MR substructures, comparing the performance of the proposed ScaMorph method with existing image registration methods. Data are represented as mean ± SEM.

representation. Among the cutting-edge approaches is VoxelMorph,[9,30] which stands out as an unsupervised end-to-end registration method, conceptualizing the registration problem as a mapping function between a fixed image and a moving image and efficiently computing the deformation field in a single step by directly evaluating this function. Another notable contribution is incorporating the cycle-consistency concept from image-to-image translation,[11] which has succeeded in the CycleGAN domain. To address the challenge of potential entrapment in local minima during the optimization process, LapIRN[31] employs a multiresolution strategy. Furthermore, LKU-Net[12] enhances the effective receptive field by integrating parallel convolutional blocks into the fundamental U-Net architecture. RDP[32] introduces a recursive deformable pyramid network for unsupervised brain MRI registration. GroupMorph[33] employs a group-wise correlation approach to decompose the deformation field into multiple subfields characterized by varying receptive fields. Despite these remarkable advancements, CNN-based methods have limitations in capturing spatial relationships between distant regions in images, restricting their applicability in medical image registration.

Transformer-based image registration methods

The adoption of the revolutionary transformer architecture, initially designed for natural language processing tasks, has extended into the computer vision domain with the emergence of the vision transformer (ViT).[13] However, applying transformers to high-resolution images and 2D structures presents certain complexities. To address these challenges and enhance the performance of vision transformers,[14] researchers have proposed several methodologies, including multi-scale architectures, light-weight convolutions, and local self-attention mechanisms.[34–36] Thanks to their self-attention mechanism, which effectively captures long-range dependencies, transformers have demonstrated their superiority over CNNs in computer vision tasks. These advancements have led to remarkable performance improvements, particularly in image registration. For instance, TransMorph[37] replaces the encoder with the Swin Transformer to enhance the modeling of spatial correspondence between input image pairs. XMorpher[38] utilizes transformers, similar to Swin Transformer, processing moving and fixed images separately and facilitating information exchange through cross-attention mechanisms. In the context of cycle-consistent image registration, Swin-VoxelMorph[39] introduces a pure transformer

Table 2. Quantitative results for brain MRI registration of the OASIS dataset from the 2021 Learn2Reg challenge task 3

Model	Dice ↑	SDlogJ ↓	HdDist95 ↓
Initial	0.5718 ± 0.0531	0.0000	3.8311
nnU-Net	0.8464 ± 0.0159	**0.0668**	1.5003
LapIRN	0.8604 ± 0.0134	0.4782	1.3748
TransMorph	0.8854 ± 0.0143	0.4953	1.2716
LKU-Net	0.8861 ± 0.0150	0.5169	1.2617
ScaMorph	**0.8886 ± 0.0148**	0.8819	1.3004

The evaluation encompasses the Dice scores for 35 cortical and subcortical brain structures, the 95th percentile of the Hausdorff distance, and the standard deviation of the logarithm of the Jacobian determinant of the displacement field for various registration methods. The validation results were directly sourced from the challenge's leaderboard. The highest scores are indicated in bold. Data are represented as mean ± SEM.

Table 3. Quantitative evaluation results for inter-modal brain MRI registration

Projects	Methods	ET		TC		WT		Average	CI	SDlogJ
		Dice ↑	HdDist95 ↓	Dice ↑	HdDist95 ↓	Dice ↑	HdDist95 ↓			
T1 → T2	VoxelMorph	0.911 ± 0.031	1.883 ± 0.345	0.891 ± 0.051	1.911 ± 0.126	0.741 ± 0.088	2.236 ± 0.000	0.848 ± 0.097	[0.741 0.911]***	**0.007 ± 0.003**
	TransMorph	0.919 ± 0.031	1.883 ± 0.345	0.904 ± 0.045	1.911 ± 0.126	0.770 ± 0.082	2.157 ± 0.111	0.864 ± 0.088	[0.770 0.919]**	0.009 ± 0.001
	LKU-Net	0.924 ± 0.032	1.794 ± 0.338	0.912 ± 0.040	1.626 ± 0.150	0.789 ± 0.073	1.626 ± 0.150	0.875 ± 0.080	[0.789 0.924]*	0.032 ± 0.007
	GroupMorph	0.925 ± 0.038	1.609 ± 0.276	0.913 ± 0.049	1.911 ± 0.126	0.793 ± 0.076	1.732 ± 0.000	0.877 ± 0.082	[0.793 0.925]*	0.041 ± 0.012
	ScaMorph	0.928 ± 0.034	1.520 ± 0.150	0.917 ± 0.044	1.821 ± 0.126	0.798 ± 0.075	1.414 ± 0.000	**0.881 ± 0.080**	[**0.798 0.928**]	0.031 ± 0.011
	VoxelMorph-diff	0.910 ± 0.031	2.068 ± 0.238	0.891 ± 0.047	2.068 ± 0.238	0.741 ± 0.083	2.157 ± 0.111	0.848 ± 0.095	[0.741 0.910]***	**0.000 ± 0.000**
	TransMorph-diff	0.921 ± 0.030	1.688 ± 0.387	0.905 ± 0.049	1.821 ± 0.126	0.771 ± 0.089	1.626 ± 0.150	0.865 ± 0.091	[0.771 0.921]***	**0.000 ± 0.000**
	LKU-Net-diff	0.924 ± 0.030	1.805 ± 0.276	0.911 ± 0.044	1.821 ± 0.126	0.785 ± 0.081	1.414 ± 0.000	0.873 ± 0.084	[0.785 0.924]**	**0.000 ± 0.000**
	GroupMorph-diff	0.924 ± 0.040	1.715 ± 0.239	0.911 ± 0.050	1.989 ± 0.206	0.790 ± 0.075	1.732 ± 0.000	0.875 ± 0.083	[0.790 0.924]*	0.044 ± 0.012
	RDP	0.927 ± 0.036	1.609 ± 0.276	0.915 ± 0.046	1.989 ± 0.206	0.796 ± 0.076	1.414 ± 0.000	0.879 ± 0.081	[0.796 0.927]*	0.030 ± 0.010
	ScaMorph-diff	0.928 ± 0.031	1.609 ± 0.276	0.918 ± 0.037	1.911 ± 0.126	0.799 ± 0.071	1.414 ± 0.000	**0.882 ± 0.077**	[**0.799 0.928**]	**0.000 ± 0.000**
T2 → T1	VoxelMorph	0.892 ± 0.031	2.157 ± 0.111	0.884 ± 0.050	2.631 ± 0.279	0.733 ± 0.085	2.378 ± 0.101	0.836 ± 0.095	[0.733 0.892]***	**0.003 ± 0.002**
	TransMorph	0.897 ± 0.029	2.229 ± 0.184	0.890 ± 0.047	2.323 ± 0.521	0.740 ± 0.087	1.989 ± 0.206	0.842 ± 0.094	[0.740 0.897]***	0.005 ± 0.001
	LKU-Net	0.892 ± 0.028	2.157 ± 0.111	0.889 ± 0.046	2.799 ± 0.404	0.744 ± 0.084	2.505 ± 0.245	0.842 ± 0.090	[0.744 0.892]***	0.015 ± 0.005
	GroupMorph	0.914 ± 0.027	1.794 ± 0.338	0.907 ± 0.043	1.989 ± 0.206	0.776 ± 0.076	1.626 ± 0.150	0.866 ± 0.082	[0.776 0.914]*	0.011 ± 0.004
	ScaMorph	0.915 ± 0.026	1.883 ± 0.345	0.909 ± 0.043	1.989 ± 0.206	0.778 ± 0.080	1.911 ± 0.126	**0.867 ± 0.084**	[**0.778 0.915**]	0.004 ± 0.003
	VoxelMorph-diff	0.910 ± 0.031	2.068 ± 0.238	0.892 ± 0.047	2.068 ± 0.238	0.741 ± 0.083	2.157 ± 0.111	0.848 ± 0.095	[0.741 0.910]***	**0.000 ± 0.000**
	TransMorph-diff	0.900 ± 0.028	2.378 ± 0.101	0.896 ± 0.047	2.323 ± 0.521	0.753 ± 0.082	2.157 ± 0.111	0.850 ± 0.089	[0.753 0.900]***	**0.000 ± 0.000**
	LKU-Net-diff	0.903 ± 0.024	2.068 ± 0.238	0.897 ± 0.044	2.505 ± 0.245	0.755 ± 0.086	2.236 ± 0.000	0.852 ± 0.089	[0.755 0.903]***	**0.000 ± 0.000**
	GroupMorph-diff	0.912 ± 0.027	.1.883 ± 0.345	0.906 ± 0.043	1.989 ± 0.206	0.772 ± 0.079	1.626 ± 0.150	0.864 ± 0.085	[0.772 0.910]*	0.013 ± 0.004
	RDP	0.916 ± 0.026	1.794 ± 0.338	0.910 ± 0.042	1.989 ± 0.206	0.779 ± 0.078	1.732 ± 0.000	0.868 ± 0.083	[0.779 0.916]*	0.005 ± 0.003
	ScaMorph-diff	0.916 ± 0.024	1.794 ± 0.338	0.910 ± 0.039	1.794 ± 0.338	0.782 ± 0.075	1.911 ± 0.126	**0.869 ± 0.080**	[**0.782 0.916**]	**0.000 ± 0.000**

Dice score and the 95th percentile of the Hausdorff distance for three tumor structures, as well as the standard deviation of the logarithm of the Jacobian determinant (SDlogJ) of the displacement field are evaluated for different methods. A 95% confidence interval (CI) was calculated for each method, with Cohen's d values indicated as follows: ∗∗∗: d ≥ 0.8, ∗∗: d ≥ 0.5, ∗: d ≥ 0.2. The bolded numbers indicate the highest scores for both non-diffeomorphic and diffeomorphic models. Data are represented as mean ± SEM.

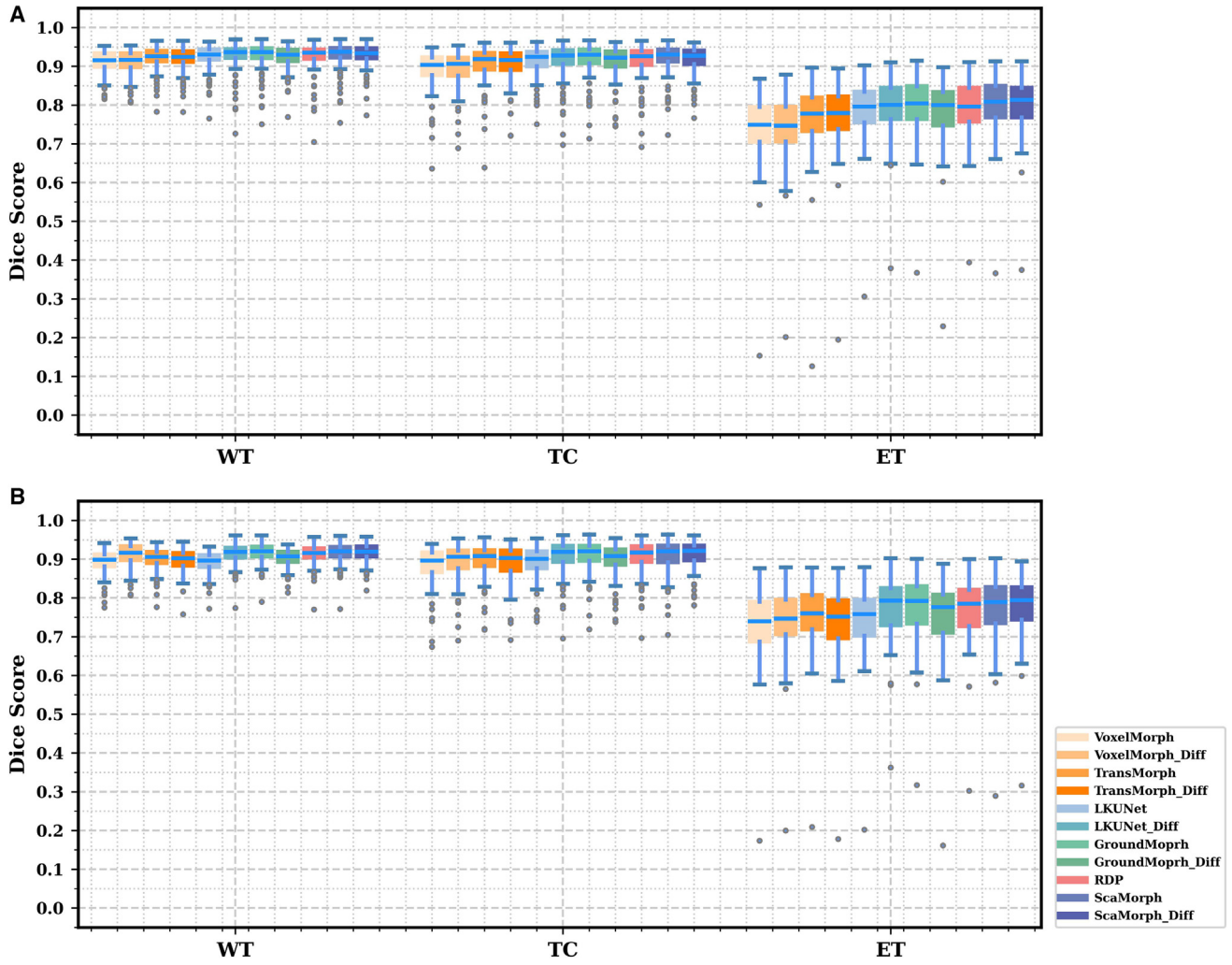


Figure 3. Quantitative comparison of the various registration methods on the inter-modal brain MR registration task

Boxplots showing Dice scores for different brain tumor substructures using the proposed ScaMorph and existing image registration methods. The figure includes both (A) T1 → T2 registration and (B) T2 → T1 registration. Data are represented as mean ± SEM.

encoder-decoder network specifically designed for this purpose. Liu et al.[40] introduce QUIZ, an innovative approach for medical image registration that leverages arbitrary voxel point of interest matching. Moreover, PC-SwinMorph[41] combines patch-level contrastive learning with Swin Transformer blocks for image registration. Despite these advancements, current hybrid networks still encounter challenges in effectively modeling long-range dependencies throughout the model, thereby hindering further performance enhancements.

RESULTS

Baseline methods

In this study, we conducted a comprehensive assessment of the ScaMorph and ScaMorph-diff model through comparisons with several established registration methods recognized for their state-of-the-art performance. Initially, we evaluated ScaMorph and ScaMorph-diff against four non-deep-learning-based methods: SyN,[7] NiftyReg [24], deedsBCV,[42] and LDDMM.[6] The hyperparameters for these methods were empirically set according to the methodology outlined by J. Chen et al.[37] Subsequently, we expanded our evaluation to include comparisons with various deep-learning-based methods, such as VoxelMorph,[9,30] VoxelMorph-diff,[10] Transmorph and its diffeomorphic variants,[37] LKU-Net and its diffeomorphic variants,[12] GroupMorph and its diffeomorphic variants[33] as well as RDP.[32] These comparisons served to elucidate the distinct capabilities and efficacy of ScaMorph in the context of deformable image registration.

Atlas-to-patient brain MRI registration

Table 1 presents the quantitative assessment findings for registering the atlas and patient images. The results demonstrate that ScaMorph achieves superior precision in image registration, with an average Dice score of 0.755, surpassing alternative techniques. Furthermore, ScaMorph-diff achieves a slightly higher average Dice score of 0.758 compared to ScaMorph.

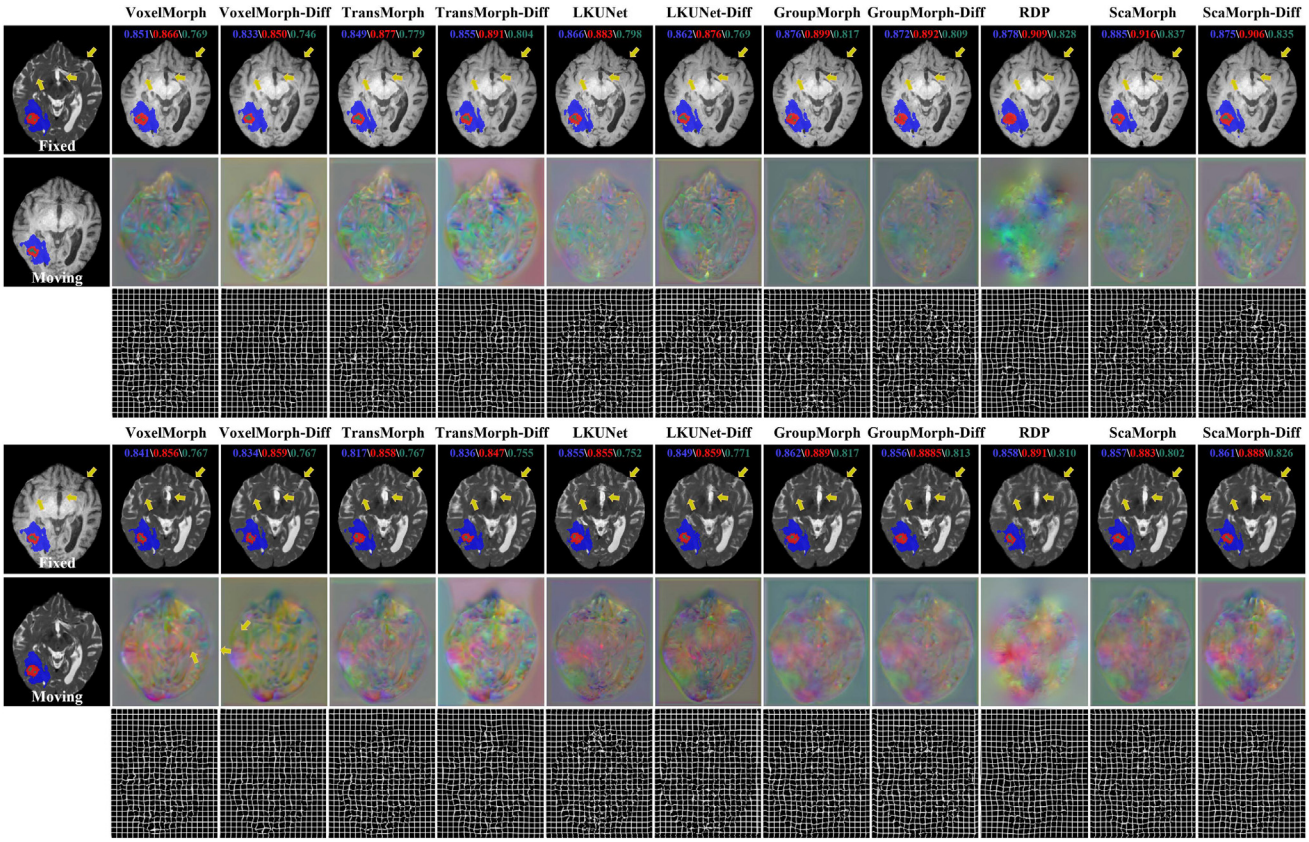


Figure 4. Qualitative results of different registration methods on the inter-modal brain MR registration task

The first row displays the deformed moving images, the second row visualizes the deformation fields, and the last row illustrates the deformed grids. Specifically, the blue, red, and green masks respectively represent the WT, ET, and TC.

It is worth noting that LKU-Net-diff outperforms other methods regarding the Jacobian determinant score. However, our proposed ScaMorph demonstrates significant advantages over the LKU-Net method in terms of Dice and Jacobian determinant scores. Moreover, a comparison is made between our approach and traditional registration methods. The table clearly illustrates that deep learning methods, with their learnable parameters, can produce more complex network models and exceptional

Table 4. Quantitative results of the inter-patient liver CT registration

Model	Dice ↑	CI	HdDist95 ↓	$\lvert J_\phi \rvert \leq 0$ ↓
VoxelMorph	0.802 ± 0.057	[0.783 0.818]***	7.760 ± 3.960	0.228 ± 0.013
TransMorph	0.812 ± 0.056	[0.794 0.829]***	7.423 ± 4.019	0.255 ± 0.010
LKU-Net	0.818 ± 0.058	[0.799 0.835]***	7.558 ± 5.976	2.275 ± 1.283
GroupMorph	0.853 ± 0.064	[0.833 0.872]***	**5.451 ± 4.840**	**0.002 ± 0.002**
ScaMorph	**0.853 ± 0.060**	[**0.834 0.872**]	5.607 ± 4.652	0.049 ± 0.015
VoxelMorph-diff	0.773 ± 0.060	[0.754 0.791]***	8.455 ± 3.979	**0.000 ± 0.000**
TransMorph-diff	0.833 ± 0.059	[0.815 0.851]***	6.683 ± 4.185	0.004 ± 0.001
LKU-Net-diff	0.831 ± 0.058	[0.813 0.848]***	7.287 ± 5.233	**0.000 ± 0.000**
GroupMorph-diff	0.852 ± 0.063	[0.831 0.870]***	5.553 ± 4.874	0.001 ± 0.001
RDP	0.857 ± 0.059	[0.838 0.876]***	5.313 ± 4.850	0.002 ± 0.001
ScaMorph-diff	**0.863 ± 0.061**	[**0.842 0.880**]	**5.084 ± 4.811**	**0.000 ± 0.000**

Dice score of the liver structures, the 95th percentile percentage of the Hausdorff distance, and the standard deviation of the logarithm of the Jacobian determinant of the displacement field are evaluated for different methods. A 95% confidence interval (CI) was calculated for each method, with Cohen's d values indicated as follows: ∗∗∗: d ≱ 0.8, ∗∗: d ≱ 0.5, ∗: d ≱ Z ≱ 0.2. The bolded numbers indicate the highest scores for both non-diffeomorphic and diffeomorphic models. Data are represented as mean ± SEM.

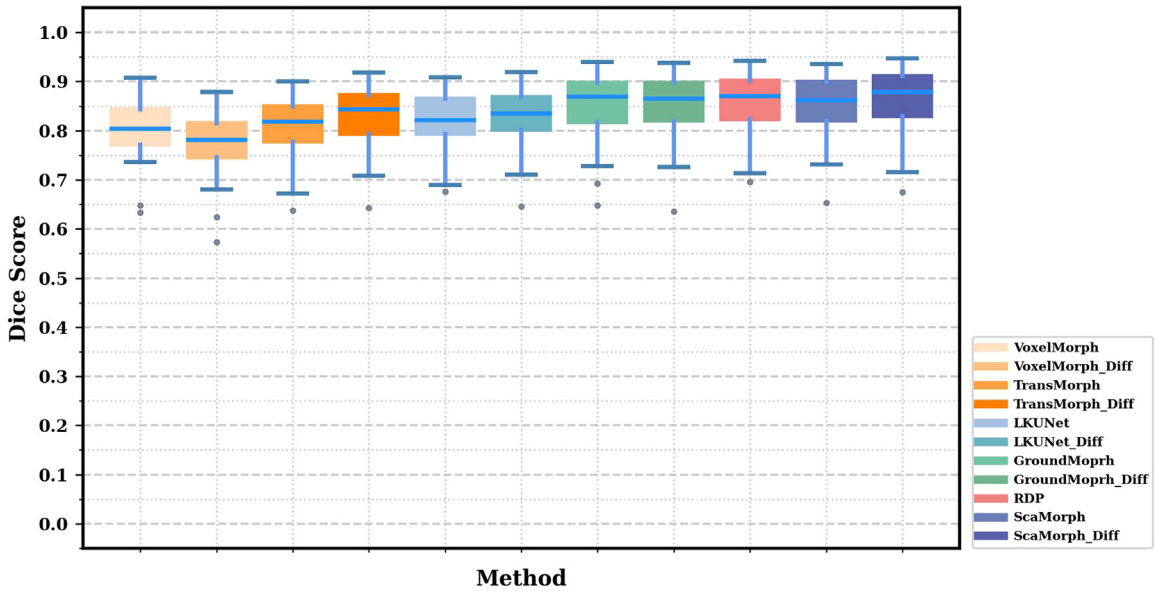


Figure 5. Quantitative evaluation results of the inter-patient liver CT registration

Boxplots depict Dice scores for different brain MR substructures, comparing the performance of the proposed ScaMorph method with existing image registration methods. Data are represented as mean ± SEM.

registration results. In our study, we utilized the Bonferroni correction (i.e., divided the *p* values by 14, the total number of paired t tests performed) to control for type I error across paired t tests, demonstrating the superior performance of ScaMorph-diff with *p* values consistently below 0.0005, thereby enhancing the reliability of our results. Furthermore, a qualitative comparison of the effects of various deep learning registration methods on MRI slice samples is conducted through graphical representations (see Figure 1). This figure demonstrates that our proposed ScaMorph and ScaMorph-diff methods yield superior qualitative results, with the deformed images resembling the fixed images more and showing enhanced alignment accuracy across different brain MRI structures. Additionally, in the depiction of deformation fields, it can be observed that both ScaMorph and ScaMorph-diff methods have advantages in the computation and visualization of deformation fields. A deformation field provides a parametric representation that elucidates the differences in shape and structure between a source image and a target image, facilitating further analysis of tissue morphological alterations or subsequent image processing tasks. As evidenced by the provided deformation field illustrations, both methods accurately capture the shape disparities and variations between the source and target images. Figure 2 offers a detailed comparison of Dice coefficients for different anatomical structures.

Inter-patient brain MRI registration

Table 2 presents the quantitative outcomes for evaluating the validation dataset in the challenge. The validation scores for different methodologies were obtained from the challenge leaderboard. With a Dice score of 0.8886, ScaMorph surpasses LKU-Net (0.8861) and TransMorph (0.8854), emerging as the highest-performing method. This achievement secured ScaMorph the esteemed fourth position. While nnU-Net achieves the highest SDlogJ score and LKU-Net outperforms others in the HdDist95 score, it is important to emphasize that ScaMorph maintains a significant advantage over its counterparts overall. By attaining the highest Dice score, ScaMorph vividly showcases its superiority in accurately aligning cerebral structures compared to LKU-Net and TransMorph, resulting in a 0.32% improvement over TransMorph. Although ScaMorph may not achieve the best scores in SDlogJ and HdDist95 metrics, the Dice score directly reflects alignment precision, the most vital metric for appraising registration performance. Thus, the comprehensive performance of ScaMorph effectively underscores its superiority over alternative methodologies. Furthermore, it is worth highlighting that ScaMorph surpasses the initial model by a substantial margin, yielding a remarkable 31.68% improvement in Dice score. This notable improvement serves as a testament to the efficacy and robust performance of ScaMorph in accurately registering brain MRI images from the esteemed OASIS dataset.

Inter-modal brain MRI registration

Table 3 presents the registration outcomes obtained from different techniques on the brain tumor dataset. The table includes forward and reverse registrations, and the assessment of performance is based on the Dice score, Hausdorff distance, and smoothness. Notably, both ScaMorph and ScaMorph-diff demonstrate competitive performance in both registration directions (T1 → T2 and T2 → T1). Specifically, ScaMorph and ScaMorph-diff achieve higher Dice scores compared to other methods for most tumor structures. In the T1 to T2 transformation, ScaMorph achieves Dice scores of 0.928, 0.918, and 0.917 for the enhancing tumor (ET), tumor core (TC), and whole tumor (WT) tumor structures, respectively, while ScaMorph-diff achieves scores of 0.928, 0.918, and 0.918. These results

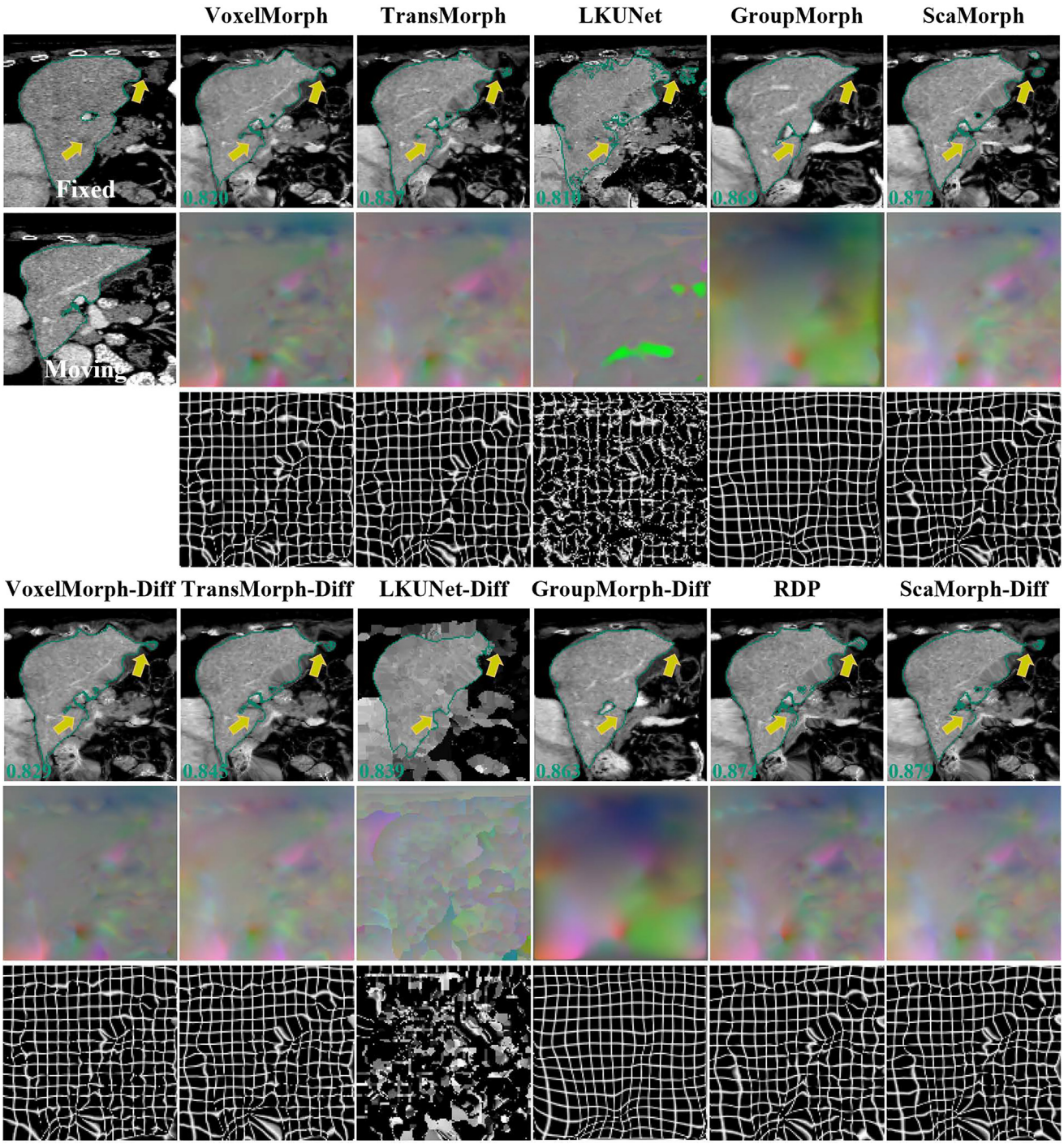


Figure 6. Qualitative results of different registration methods on the inter-patient liver CT registration task

The first row displays the deformed moving images, the second row visualizes the deformation fields, and the last row illustrates the deformed grids. Specifically, the green contours respectively represent the liver.

indicate that ScaMorph and ScaMorph-diff accurately preserve and align the similarity between brain structures. Additionally, ScaMorph and ScaMorph-diff consistently demonstrate the lowest scores in terms of the 95th percentile of the Hausdorff distance. In the T1 to T2 transformation, ScaMorph and ScaMorph-diff exhibit HdDist95 scores of 1.520 and 1.609, respectively, while other methods have scores of 2.068 and above. This indicates that ScaMorph and ScaMorph-diff excel in minimizing the dissimilarities between brain structures. Furthermore, ScaMorph and ScaMorph-diff also showcase

Table 5. Quantitative evaluation results for inter-modal abdomen MR-CT registration

Model	Dice ↑	CI	HdDist95 ↓	$\lvert J_\phi \rvert \leq 0$ ↓
VoxelMorph	0.717 ± 0.190	[0.614 0.802]***	0.554 ± 0.271	0.147 ± 0.013
TransMorph	0.725 ± 0.181	[0.615 0.796]***	0.555 ± 0.270	0.109 ± 0.017
LKU-Net	0.773 ± 0.186	[0.658 0.846]***	0.596 ± 0.290	0.141 ± 0.028
GroupMorph	0.787 ± 0.162	[0.692 ± 0.851]***	0.640 ± 0.257	0.127 ± 0.016
ScaMorph	**0.812 ± 0.135**	[**0.735 0.866**]	**0.692 ± 0.219**	**0.106 ± 0.019**
VoxelMorph-diff	0.725 ± 0.181	[0.615 0.796]***	0.555 ± 0.270	**0.000 ± 0.000**
TransMorph-diff	0.762 ± 0.184	[0.655 0.839]***	0.595 ± 0.292	0.002 ± 0.001
LKU-Net-diff	0.785 ± 0.173	[0.673 0.857]***	0.633 ± 0.278	**0.000 ± 0.000**
GroupMorph-diff	0.779 ± 0.164	[0.676 0.851]***	0.619 ± 0.243	0.002 ± 0.001
RDP	0.793 ± 0.167	[0.710 0.859]**	0.661 ± 0.257	0.006 ± 0.003
ScaMorph-diff	**0.812 ± 0.140**	[**0.726 0.868**]	**0.680 ± 0.221**	**0.000 ± 0.000**

Dice score and the 95th percentile of the Hausdorff distance for three tumor structures, as well as the standard deviation of the logarithm of the Jacobian determinant (SDlogJ) of the displacement field are evaluated for different methods. A 95% confidence interval (CI) was calculated for each method, with Cohen's d values indicated as follows: ***: d ≱ 0.8, **: d ≱ 0.5, ≱: d ≱ 0.2. The bolded numbers indicate the highest scores for both non-diffeomorphic and diffeomorphic models. Data are represented as mean ± SEM.

commendable performance in terms of the standard deviation of the logarithm of the Jacobian determinant (SDlogJ) of the displacement field. Although their scores are not consistently the lowest in certain cases, they remain comparable to other methods in most situations. Paired t-tests with Bonferroni correction resulted in p values of $p \ll 0.0005$ when comparing the most superior model, ScaMorph-diff, with all other methods. This statistical evidence suggests that the proposed approach significantly outperforms the comparative registration methods and network architectures, accurately preserving and aligning the similarity between brain structures while mitigating the differences. This is of great importance for brain image analysis and research. Figure 3 provides a quantitative comparison of inter-modal brain MRI registration, showing that both ScaMorph and its diffeomorphic variants achieve superior Dice metric scores for the majority of tumor structures in both registration directions.

Figure 4 depicts a qualitative comparison between our proposed ScaMorph and ScaMorph-diff techniques and other alignment approaches on MRI slice samples. It can be observed that ScaMorph and ScaMorph-diff perform admirably in terms of alignment accuracy for these three tumor structures, showcasing exceptional alignment accuracy. This implies that both methods excel in maintaining structural and shape accuracy when mapping different types of tissues onto the reference image. The visualization of the deformation fields (rows 2 and 3) illustrates that our proposed registration model can generate relatively smooth deformation fields and perform well in terms of edge preservation and detail recovery.

Inter-patient liver CT registration

Table 4 presents the quantitative assessment outcomes of the registration process applied to the patients' CT images. Similar to the findings discussed in the preceding sections, ScaMorph and ScaMorph-diff demonstrated superior precision in the registration results. In terms of the Dice metric, ScaMorph achieved a score of 0.853 ± 0.060, indicating the highest performance compared to the other non-diffeomorphic models (VoxelMorph, TransMorph, and LKU-Net). Additionally, ScaMorph achieved a score of 5.607 ± 4.652 on the HD95 metric, suggesting a closer alignment to the actual scenario. Moreover, ScaMorph exhibited a low level of 0.049 ± 0.015 on the percentage of metrics with non-normal comparable determinants (folded pixels), further supporting its effective preservation of topological properties. ScaMorph-diff obtained scores of 0.863 ± 0.061 on the Dice metric and 5.084 ± 4.811 on the HD95 metric, demonstrating a significant improvement over other diffeomorphic registration methods. The percentage of non-orthonormal comparable determinants achieved by ScaMorph-diff also remained consistently low (less than 0.0001), indicating the preservation of its topological properties. Figure 5 presents a quantitative comparison of inter-patient liver CT registration. The figure clearly illustrates that both ScaMorph and its diffeomorphic variants achieved superior scores in the Dice metric. Paired t tests and Bonferroni correction performed between ScaMorph and all other methods produced $p \ll 0.0005$. The top-right panel of Figure 6 exhibits the qualitative outcomes of ScaMorph and other registration techniques on a sample slice for inter-patient liver CT registration.

Inter-modal abdomen MRI-CT registration

Table 5 presents the quantitative evaluation results of the registration process applied to abdomen MRI-CT images. The findings reveal that both ScaMorph and ScaMorph-diff achieved the highest Dice scores of 0.812 ± 0.135 and 0.812 ± 0.140, respectively, outperforming other non-diffeomorphic and diffeomorphic models. Additionally, the HD95 metric scores for ScaMorph and ScaMorph-diff were 0.692 ± 0.219 and 0.680 ± 0.221, respectively, representing the lowest values recorded among all evaluated methods. Notably, the percentage of non-orthonormal comparable determinants for ScaMorph-diff remained consistently low (less than 0.0001), indicating effective preservation of topological properties. Figure 7 offers a quantitative comparison of

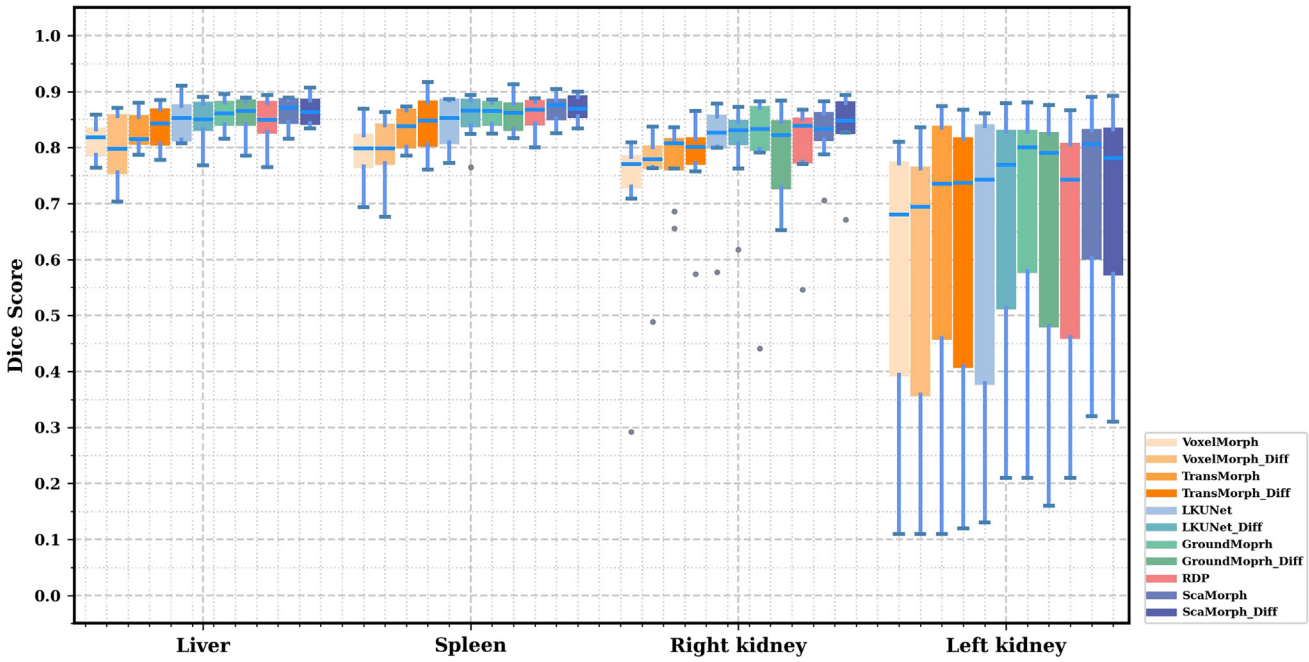


Figure 7. Quantitative evaluation results of the inter-modal abdomen MR-CT registration

Boxplots depict Dice scores for different Abdomen substructures, comparing the performance of the proposed ScaMorph method with existing image registration methods. Data are represented as mean ± SEM.

inter-modal abdomen MRI-CT registration, demonstrating that both ScaMorph and its diffeomorphic variants achieved superior Dice metric scores across the majority of abdominal structures. Furthermore, Figure 8 illustrates the qualitative outcomes of ScaMorph in comparison with other registration techniques on a sample slice for inter-modal abdomen MRI-CT registration. Despite the substantial and complex deformations between the fixed and moving images, our method effectively aligns corresponding anatomical structures, yielding accurate registration results.

Ablation studies

The effectiveness of the proposed ScaMorph was further investigated through ablation studies to evaluate the contributions of the self-attention mechanism and residual blocks. Table 6 presents the quantitative outcomes of the ablation studies, demonstrating that the proposed ScaMorph outperforms other variants in terms of Dice scores and SDlogJ scores. This indicates the essential nature of both the self-attention mechanism and residual blocks for the proposed ScaMorph. The self-attention mechanism is capable of capturing long-range dependencies and modeling complex features, while the residual blocks facilitate the flow of information and enhance the learning capacity of the network. The combination of these two components enables the proposed ScaMorph to achieve superior performance in terms of registration accuracy.

Convergence and complexity

We assessed the influence of model complexity on registration performance by analyzing the computational complexity and parameter counts of various deep learning architectures, as illustrated in Figure 9. Complexity calculations were performed using an input image size of 160 × 192 × 224, representative of brain MRI scans. ScaMorph exhibits a commendable equilibrium between accuracy and computational efficiency. With a floating point operations (FLOPs) count of approximately 845.55 million, it significantly surpasses the efficiency of methods such as RDP, which has an FLOP count of 4046.21 million, while also outperforming models like TransMorph and LKUNet in terms of parameter efficiency. Importantly, ScaMorph's parameter count of 8.82 million positions it advantageously against TransMorph, which exceeds 46 million parameters. Table 7 presents a comparison of training time in minutes per epoch (min/epoch) and inference time in seconds per image (sec/image) among the methodologies employed in this study. It is noteworthy that the SyN, NiftyReg, and deedsBCV packages are CPU-based, whereas LDDMM and the deep learning methods are GPU-based. ScaMorph's training time is recorded at 6.159 min per epoch, while its inference time is 0.513 s per image. Compared to other models listed in the table, ScaMorph's inference time is relatively advantageous, particularly when contrasted with methods such as TransMorph (0.558 s/image) and GroupMorph (0.507 s/image), which, despite having comparable computational profiles, do not exhibit a significant advantage over ScaMorph. In summary, ScaMorph not only achieves superior accuracy metrics but also preserves a favorable computational profile, thus rendering it suitable for real-time medical image processing applications. This analysis highlights the critical need to balance performance with computational demands, particularly within clinical settings where timely results are paramount.

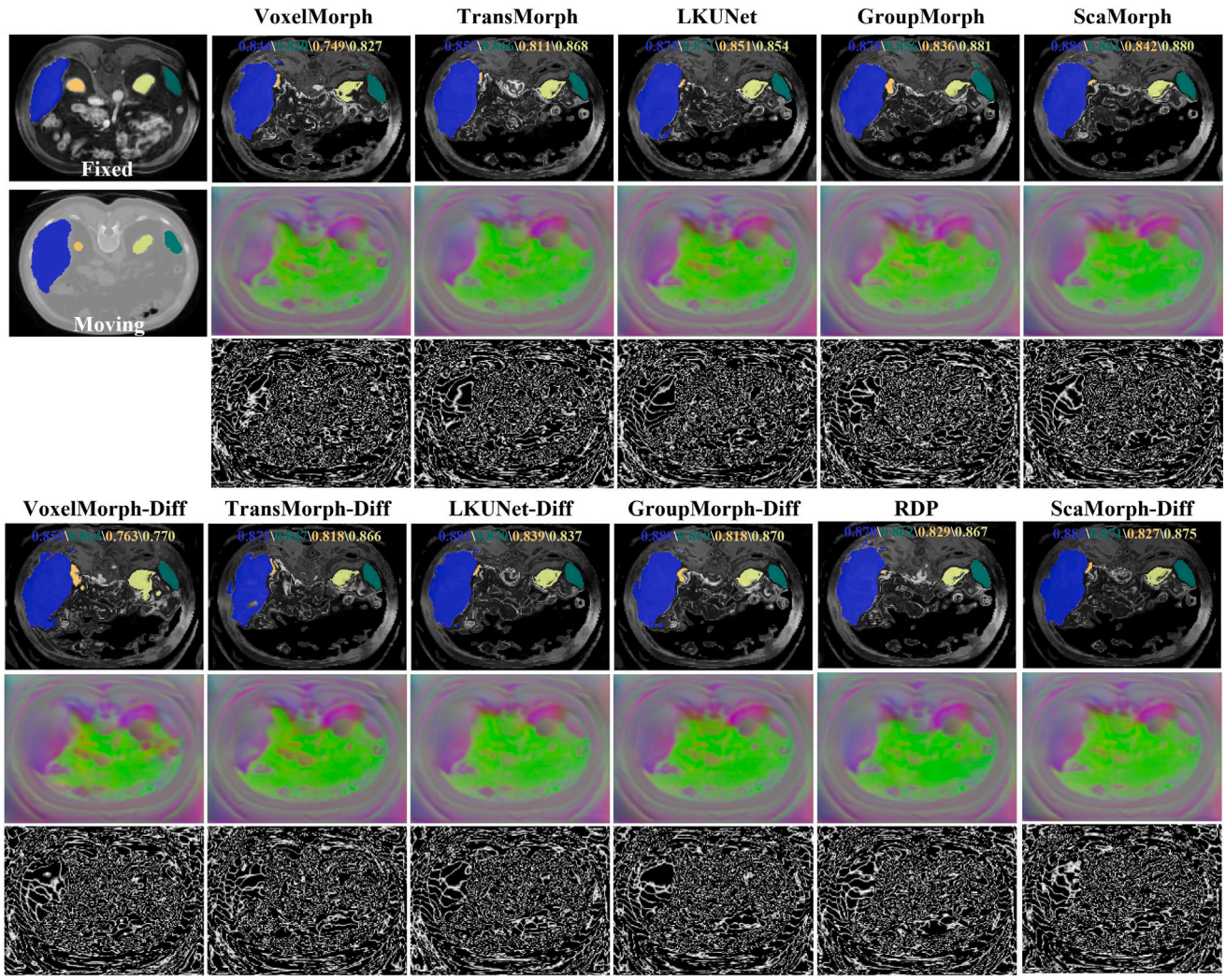


Figure 8. Qualitative results of different registration methods on the inter-modal abdomen MR-CT registration task

The first row displays the deformed moving images, the second row visualizes the deformation fields, and the last row illustrates the deformed grids. Specifically, the blue, yellow, orange and green masks respectively represent the liver, left kidney, right kidney and spleen.

DISCUSSION

In this study, we introduce an innovative unsupervised deformable image registration method based on scale-aware context aggregation (ScaMorph), which demonstrates superior accuracy and overall performance in brain and liver image registration. Specifically, we developed the MMC module and the MCF module, which aid in capturing complex features and long-range dependencies, enhancing volumetric overlap between anatomical structures, and reducing the number of folded voxels during the registration process.

The proposed ScaMorph method was compared with existing state-of-the-art registration methods on three MRI datasets from different registration tasks and modalities, including a 3D liver dataset and an inter-modal abdomen MRI-CT dataset. Through comparisons in atlas-to-patient brain MRI, inter-patient brain MRI, inter-modal brain MRI, inter-patient liver CT registration, and inter-modal abdomen MRI-CT registration tasks, ScaMorph demonstrated superior accuracy in anatomical structure matching and overall registration performance. The versatility of ScaMorph extends beyond MRI and liver image registration, suggesting its applicability to a broader range of medical imaging modalities. The potential applications of ScaMorph encompass various medical domains, including disease diagnosis and progression monitoring, surgical planning and navigation, and potentially other areas requiring precise image registration.

Additionally, the ScaMorph-diff, a diffeomorphic variant of the proposed method, was also evaluated in the aforementioned experiments. The experimental results indicate that this method surpasses existing methods in both qualitative and quantitative performance, particularly in balancing enhanced visual quality of images with the preservation of tissue structure details. Even across different modalities, this method exhibited better

Table 6. Results of ablation studies

Model	Dice ↑	$\lvert J_\phi\rvert \leq 0$ ↓
ScaMorph w/o SCA & Res	0.751 ± 0.131	1.625 ± 0.363
ScaMorph w/o SCA	0.752 ± 0.131	1.539 ± 0.325
ScaMorph w/o Res	0.753 ± 0.131	1.448 ± 0.356
ScaMorph	**0.755 ± 0.132**	**1.400 ± 0.329**

The highest scores are indicated in bold. Data are represented as mean ± SEM.

performance, highlighting its robustness, generalization capability, and detail capturing ability. Furthermore, the proposed method is suitable for deep learning-based multimodal fusion and can be extended to other medical imaging modalities, such as image denoising and brain atlas construction.

In our experiments, ScaMorph demonstrated a marked enhancement in alignment accuracy compared to TransMorph across multiple datasets, particularly under conditions of pronounced anatomical variability. The findings indicate that ScaMorph is superior in managing complex alignment tasks, as evidenced by the significantly improved alignment accuracy observed in the OASIS experiments. This improvement is attributable to the introduction of a effective SCA mechanism, which augments the model's capacity to concentrate on pertinent spatial features during the registration process. Consequently, ScaMorph achieves a more nuanced comprehension of anatomical structures, thereby enhancing alignment accuracy in challenging scenarios.

This study elucidates the substantial clinical applications of ScaMorph through various registration experiments, including atlas-to-patient brain MRI registration, inter-patient brain MRI registration, inter-modal T1 and T2 brain tumor registration, inter-patient liver CT registration, and inter-modal abdomen MRI-CT registration. For instance, in atlas-to-patient brain MRI registration, ScaMorph significantly improves the alignment of individual patient scans with a standardized brain template, thereby enhancing diagnostic accuracy in the identification of abnormalities. The inter-patient registration experiment further highlights its proficiency in accurately aligning images from disparate patients, a critical factor for comparative studies and personalized treatment strategies. In the context of T1 and T2 brain tumor registration, ScaMorph facilitates precise tracking of tumor evolution over time, which is instrumental in monitoring treatment responses and disease progression. Additionally, the liver CT registration underscores its utility in surgical planning, where accurate image alignment can inform interventions and mitigate procedural risks. Lastly, the Abdomen MRI-CT registration experiment accentuates ScaMorph's versatility across imaging modalities, enabling comprehensive evaluations of abdominal pathologies. By integrating these applications, ScaMorph not only enhances technical performance but also optimizes clinical workflows, ultimately improving decision-making and patient outcomes.

Limitations of the study

Despite the promising advancements offered by ScaMorph, we acknowledge certain limitations that beckon further exploration. The current study's focus on brain and liver organ registrations,

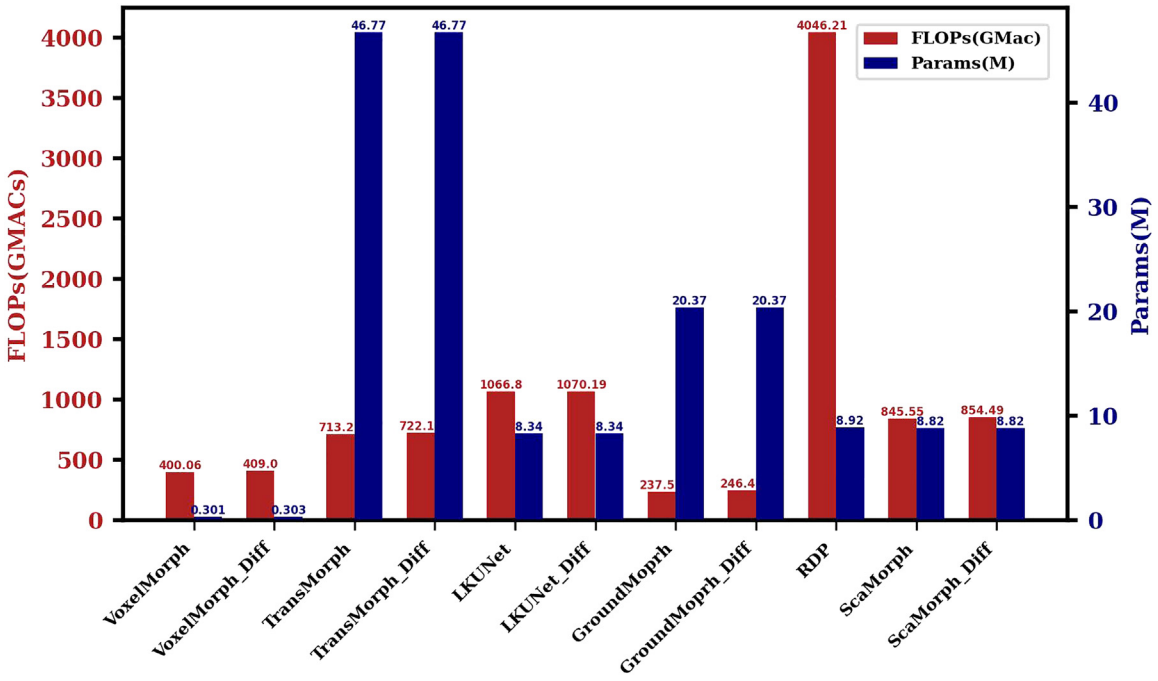


Figure 9. Convergence analysis of different registration methods

The number of parameters are in units of millions of parameters. Model computational complexity comparisons represented in GMACs. Greater values imply a greater degree of computational complexity. These values were obtained using an input image of size **160 × 192 × 224**.

Table 7. Average training and inference times for different methods used in this study

Model	Training Time(min/epoch)	Inference Time(sec/image)
SyN	–	165.356
NiftyReg	–	28.451
LDDMM	–	53.756
deedsBCV	–	28.842
VoxelMorph	3.244	0.261
TransMorph	7.198	0.558
LKU-Net	6.862	0.481
GroupMorph	10.161	0.507
ScaMorph	6.159	0.513
VoxelMorph-diff	3.452	0.294
TransMorph-diff	8.098	0.593
LKU-Net-diff	7.949	0.562
GroupMorph-diff	16.369	0.589
RDP	13.057	0.376
ScaMorph-diff	7.713	0.544

Note that SyN, NiftyReg, and deedsBCV were executed on CPUs, while LDDMM and learning-based methods utilized GPUs. Inference times are averaged over 30 repeated runs.

for instance, restricts the generalizability of our findings. Future research should aim to extend the model's capabilities to organs experiencing larger deformations, such as the lungs and breasts, to validate its efficacy across a broader spectrum of medical scenarios. Moreover, the limitation to T1 and T2 image registrations points to an opportunity for expanding the model's utility to encompass registrations between MRI and CT scans or other imaging modalities. Such expansions could significantly enhance ScaMorph's applicability, making it an invaluable tool across a wider range of medical imaging contexts. Additionally, the prospective integration of segmentation and registration processes presents an exciting avenue for research, promising to simultaneously enhance the accuracy of both tasks. This could lead to more comprehensive and efficient approaches to medical image analysis and interpretation, further solidifying the role of advanced models like ScaMorph in revolutionizing medical diagnostics and treatment planning.

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources should be directed to and fulfilled by the lead contact, Yang Gao (yanggao@buaa.edu.cn).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- This paper analyzes existing, publicly available data. These accession numbers for the datasets are listed in the key resources table.
- All original code has been deposited at Github (https://github.com/Liuyuchen0224/ScaMorph) and is publicly available as of the date of publication.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

ACKNOWLEDGMENTS

This work was supported by the Joint Funds of the National Natural Science Foundation of China (no. U23A20434) and Innovation Program for Quantum Science and Technology, Hefei National Laboratory, Hefei 230088, China (no. 2021ZD0300500/2021ZD03005 03).

AUTHOR CONTRIBUTIONS

Conceptualization, Y.L. and D.W.; methodology, Y.L. and D.W.; investigation, Y.L. and L.W.; writing—original draft, Y.L. and L.W.; writing—review and editing, Y.L. and L.W.; funding acquisition, X.L., Y.G., and D.W.; resources, X.L. and Y.G.; supervision, X.L. and Y.G.

DECLARATION OF INTERESTS

The authors declare no competing interests.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- METHOD DETAILS
  - Datasets and preprocessing
  - Implementation
  - Evaluation metrics
  - Overall architecture
  - Deformable registration network
  - Scale-aware context aggregation
  - Loss function
  - Probabilistic variants
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2024.111734.

Received: July 15, 2024
Revised: September 24, 2024
Accepted: December 30, 2024
Published: January 3, 2025

REFERENCES

1. Sotiras, A., Davatzikos, C., and Paragios, N. (2013). Deformable Medical Image Registration: A Survey. IEEE Trans. Med. Imag. *32*, 1153–1190.
2. Schnabel, J.A., Heinrich, M.P., Papież, B.W., and Brady, S.J.M. (2016). Advances and challenges in deformable image registration: From image fusion to complex motion modelling. Med. Image Anal. *33*, 145–148.
3. Liu, R., Li, Z., Fan, X., Zhao, C., Huang, H., and Luo, Z. (2022). Learning Deformable Image Registration From Optimization: Perspective, Modules, Bilevel Training and Beyond. IEEE Trans. Pattern Anal. Mach. Intell. *44*, 7688–7704.
4. Chen, X., Wang, X., Zhang, K., Fung, K.-M., Thai, T.C., Moore, K., Mannel, R.S., Liu, H., Zheng, B., and Qiu, Y. (2022). Recent advances and clinical applications of deep learning in medical image analysis. Med. Image Anal. *79*, 102444.
5. Gaens, T., Maes, F., Vandermeulen, D., and Suetens, P. (1998). Non-rigid multimodal image registration using mutual information. In Medical Image

Computing and Computer-Assisted Intervention — MICCAI'98, Lecture Notes in Computer Science (Springer), pp. 1099–1106.

6. Beg, M.F., Miller, M.I., Trouvé, A., and Younes, L. (2005). Computing Large Deformation Metric Mappings via Geodesic Flows of Diffeomorphisms. Int. J. Comput. Vis. *61*, 139–157.
7. Avants, B.B., Epstein, C.L., Grossman, M., and Gee, J.C. (2008). Symmetric diffeomorphic image registration with cross-correlation: Evaluating automated labeling of elderly and neurodegenerative brain. Med. Image Anal. *12*, 26–41.
8. Klein, S., Staring, M., Murphy, K., Viergever, M.A., and Pluim, J.P.W. (2010). Elastix: A Toolbox for Intensity-Based Medical Image Registration. IEEE Trans. Med. Imag. *29*, 196–205.
9. Balakrishnan, G., Zhao, A., Sabuncu, M.R., Guttag, J., and Dalca, A.V. (2019). VoxelMorph: A Learning Framework for Deformable Medical Image Registration. IEEE Trans. Med. Imag. *38*, 1788–1800.
10. Dalca, A.V., Balakrishnan, G., Guttag, J., and Sabuncu, M.R. (2019). Unsupervised learning of probabilistic diffeomorphic registration for images and surfaces. Med. Image Anal. *57*, 226–236.
11. Kim, B., Kim, D.H., Park, S.H., Kim, J., Lee, J.-G., and Ye, J.C. (2021). CycleMorph: Cycle consistent unsupervised deformable image registration. Med. Image Anal. *71*, 102036.
12. Jia, X., Bartlett, J., Zhang, T., Lu, W., Qiu, Z., and Duan, J. (2022). U-Net vs Transformer: Is U-Net Outdated in Medical Image Registration? In Machine Learning in Medical Imaging, Lecture Notes in Computer Science (Springer Nature Switzerland), pp. 151–160.
13. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is All you Need. Adv. Neural Inf. Process. Syst. *30*, 6000–6010.
14. Liu, Z., Ning, J., Cao, Y., Wei, Y., Zhang, Z., Lin, S., and Hu, H. (2022). Video Swin Transformer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3202–3211.
15. Hou, Q., Lu, C.-Z., Cheng, M.-M., and Feng, J. (2022). Conv2Former: A Simple Transformer-Style ConvNet for Visual Recognition. Preprint at arXiv. https://doi.org/10.48550/arXiv.2211.11943.
16. Yang, J., Li, C., Dai, X., and Gao, J. (2022). Focal Modulation Networks. Adv. Neural Inf. Process. Syst. *35*, 4203–4217.
17. Guo, M.-H., Lu, C.-Z., Liu, Z.-N., Cheng, M.-M., and Hu, S.-M. (2023). Visual attention network. Comput. Vis. Media (Beijing). *9*, 733–752.
18. Lin, W., Wu, Z., Chen, J., Huang, J., and Jin, L. (2023). Scale-Aware Modulation Meet Transformer. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 6015–6026.
19. Wolberg, G., and Zokai, S. (2000). Robust image registration using log-polar transform. Proceedings 2000 International Conference on Image Processing (Cat. No.00CH37101) *1*, 493–496.
20. Schlachter, M., Fechter, T., Jurisic, M., Schimek-Jasch, T., Oehlke, O., Adebahr, S., Birkfellner, W., Nestle, U., and Bühler, K. (2016). Visualization of Deformable Image Registration Quality Using Local Image Dissimilarity. IEEE Trans. Med. Imag. *35*, 2319–2328.
21. Nie, Z., and Yang, X. (2019). Deformable Image Registration Using Functions of Bounded Deformation. IEEE Trans. Med. Imag. *38*, 1488–1500.
22. Maes, F., Vandermeulen, D., and Suetens, P. (2003). Medical image registration using mutual information. Proc. IEEE *91*, 1699–1722.
23. Ashburner, J. (2007). A fast diffeomorphic image registration algorithm. Neuroimage *38*, 95–113.
24. Modat, M., Ridgway, G.R., Taylor, Z.A., Lehmann, M., Barnes, J., Hawkes, D.J., Fox, N.C., and Ourselin, S. (2010). Fast free-form deformation using graphics processing units. Comput. Methods Progr. Biomed. *98*, 278–284.
25. Thirion, J.P. (1998). Image matching as a diffusion process: An analogy with maxwell's demons. Med. Image Anal. *2*, 243–260.
26. Vercauteren, T., Pennec, X., Perchant, A., and Ayache, N. (2009). Diffeomorphic demons: Efficient non-parametric image registration. Neuroimage *45*, S61–S72.
27. Sokooti, H., de Vos, B., Berendsen, F., Lelieveldt, B.P.F., Išgum, I., and Staring, M. (2017). Nonrigid Image Registration Using Multi-scale 3D Convolutional Neural Networks. In Medical Image Computing and Computer Assisted Intervention - MICCAI 2017, Lecture Notes in Computer Science (Springer International Publishing), pp. 232–239.
28. Yang, X., Kwitt, R., Styner, M., and Niethammer, M. (2017). Quicksilver: Fast predictive image registration – A deep learning approach. Neuroimage *158*, 378–396.
29. Xiao, H., Teng, X., Liu, C., Li, T., Ren, G., Yang, R., Shen, D., and Cai, J. (2021). A review of deep learning-based three-dimensional medical image registration methods. Quant. Imaging Med. Surg. *11*, 4895–4916.
30. Balakrishnan, G., Zhao, A., Sabuncu, M.R., Guttag, J., and Dalca, A.V. (2018). An Unsupervised Learning Model for Deformable Medical Image Registration. In IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9252–9260.
31. Mok, T.C.W., and Chung, A.C.S. (2020). Large Deformation Diffeomorphic Image Registration with Laplacian Pyramid Networks. In Medical Image Computing and Computer Assisted Intervention – MICCAI 2020, Lecture Notes in Computer Science (Springer International Publishing), pp. 211–221.
32. Wang, H., Ni, D., and Wang, Y. (2024). Recursive Deformable Pyramid Network for Unsupervised Medical Image Registration. IEEE Trans. Med. Imag. *43*, 2229–2240. https://doi.org/10.1109/TMI.2024.3362968.
33. Tan, Z., Zhang, L., Lv, Y., Ma, Y., and Lu, H. (2024). GroupMorph: Medical Image Registration via Grouping Network with Contextual Fusion. IEEE Trans. Med. Imag. *43*, 3807–3819. https://doi.org/10.1109/TMI.2024.3400603.
34. Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., and Wang, M. (2023). Swin-Unet: Unet-Like Pure Transformer for Medical Image Segmentation. In Computer Vision – ECCV 2022 Workshops, Lecture Notes in Computer Science (Springer Nature Switzerland), pp. 205–218.
35. Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A.L., and Zhou, Y. (2021). TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation. Preprint at arXiv, 04306. https://doi.org/10.48550/arXiv.2102.04306.
36. Shamshad, F., Khan, S., Zamir, S.W., Khan, M.H., Hayat, M., Khan, F.S., and Fu, H. (2023). Transformers in medical imaging: A survey. Med. Image Anal. *88*, 102802.
37. Chen, J., Frey, E.C., He, Y., Segars, W.P., Li, Y., and Du, Y. (2022). TransMorph: Transformer for unsupervised medical image registration. Med. Image Anal. *82*, 102615.
38. Shi, J., He, Y., Kong, Y., Coatrieux, J.-L., Shu, H., Yang, G., and Li, S. (2022). XMorpher: Full Transformer for Deformable Medical Image Registration via Cross Attention. In Medical Image Computing and Computer Assisted Intervention – MICCAI 2022, Lecture Notes in Computer Science (Springer Nature Switzerland), pp. 217–226.
39. Zhu, Y., and Lu, S. (2022). Swin-VoxelMorph: A Symmetric Unsupervised Learning Model for Deformable Medical Image Registration Using Swin Transformer. In Medical Image Computing and Computer Assisted Intervention – MICCAI 2022, Lecture Notes in Computer Science (Springer Nature Switzerland), pp. 78–87.
40. Liu, L., Fan, X., Liu, H., Zhang, C., Kong, W., Dai, J., Jiang, Y., Xie, Y., and Liang, X. (2024). QUIZ: An arbitrary volumetric point matching method for medical image registration. Comput. Med. Imag. Graph. *112*, 102336. https://doi.org/10.1016/j.compmedimag.2024.102336.
41. Yin, W., Sonke, J.-J., and Gavves, E. (2023). PC-Reg: A pyramidal prediction–correction approach for large deformation image registration. Med. Image Anal. *90*, 102978.
42. Heinrich, M.P., Maier, O., and Handels, H. (2015). Multi-modal Multi-Atlas Segmentation using Discrete Optimisation and Self-Similarities. VISCERAL Challenge@ ISBI *1390*, 27.
43. Marcus, D.S., Wang, T.H., Parker, J., Csernansky, J.G., Morris, J.C., and Buckner, R.L. (2007). Open Access Series of Imaging Studies (OASIS): Cross-sectional MRI Data in Young, Middle Aged, Nondemented, and Demented Older Adults. J. Cognit. Neurosci. *19*, 1498–1507.

44. Baid, U., Ghodasara, S., Mohan, S., Bilello, M., Calabrese, E., Colak, E., Farahani, K., Kalpathy-Cramer, J., Kitamura, F.C., Pati, S., and Prevedello, L.M. (2021). The RSNA-ASNR-MICCAI BraTS 2021 Benchmark on Brain Tumor Segmentation and Radiogenomic Classification. Preprint at arXiv. https://doi.org/10.48550/arXiv.2107.02314.

45. Bilic, P., Christ, P., Li, H.B., Vorontsov, E., Ben-Cohen, A., Kaissis, G., Szeskin, A., Jacobs, C., Mamani, G.E.H., Chartrand, G., et al. (2023). The Liver Tumor Segmentation Benchmark (LiTS). Med. Image Anal. *84*, 102680.

46. Clark, K., Vendt, B., Smith, K., Freymann, J., Kirby, J., Koppel, P., Moore, S., Phillips, S., Maffitt, D., Pringle, M., et al. (2013). The cancer imaging archive (tcia): maintaining and operating a public information repository. J. Digit. Imag. *26*, 1045–1057.

47. Zhao, S., Lau, T., Luo, J., Chang, E.I.-C., and Xu, Y. (2020). Unsupervised 3D End-to-End Medical Image Registration with Volume Tweening Network. IEEE J. Biomed. Health Inform. *24*, 1394–1404.

48. Fischl, B. (2012). FreeSurfer. Neuroimage *62*, 774–781.

49. Hering, A., Hansen, L., Mok, T.C.W., Chung, A.C.S., Siebert, H., Hager, S., Lange, A., Kuckertz, S., Heldmann, S., Shao, W., et al. (2023). Learn2Reg: Comprehensive Multi-Task Medical Image Registration Challenge, Dataset and Evaluation in the Era of Deep Learning. IEEE Trans. Med. Imag. *42*, 697–712.

50. Kong, L., Qi, X.S., Shen, Q., Wang, J., Zhang, J., Hu, Y., and Zhou, Q. (2023). Indescribable Multi-Modal Spatial Evaluator. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9853–9862.

51. Dice, L.R. (1945). Measures of the Amount of Ecologic Association Between Species. Ecology *26*, 297–302.

52. Huttenlocher, D., Klanderman, G., and Rucklidge, W. (1993). Comparing images using the Hausdorff distance. IEEE Trans. Pattern Anal. Mach. Intell. *15*, 850–863.

53. Jaderberg, M., Simonyan, K., Zisserman, A., and Kavukcuoglu, K. (2015). Spatial Transformer Networks. Adv. Neural Inf. Process. Syst. *28*.

54. He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778.

55. Mok, T.C.W., and Chung, A.C.S. (2020). Fast Symmetric Diffeomorphic Image Registration with Convolutional Neural Networks. In IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4644–4653.

56. Kingma, D.P., and Welling, M. (2022). Auto-Encoding Variational Bayes. Preprint at arXiv. https://doi.org/10.48550/arXiv.1312.6114.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
RRID:SCR_005839	Biomedical Image Analysis Group	https://brain-development.org/ixi-dataset/
RRID:SCR_007385	Marcus et al.[43]	https://learn2reg.grand-challenge.org/Learn2Reg2021/
RRID:SCR_016214	Baid et al.[44]	https://www.med.upenn.edu/cbica/brats2021/
LiTS	Bilic et al.[45]	https://competitions.codalab.org/competitions/17094
AbdomenMRCT	Clark et al.[46]	https://learn2reg.grand-challenge.org/Learn2Reg2021/
Software and algorithms		
RRID:SCR_008394	Python Software Foundation	https://www.python.org/
RRID:SCR_018536	PyTorch Foundation	https://pytorch.org/
ScaMorph	This paper	https://github.com/Liuyuchen0224/ScaMorph

METHOD DETAILS

Datasets and preprocessing

To comprehensively validate the proposed methodology, a series of experiments were conducted following the methodology outlined in ref. [37] and [47]. Five distinct datasets, comprising over 1500 pairs of images, were utilized. Each dataset played a crucial role in evaluating the robustness and efficacy of the proposed methodology. Below are the details of each dataset and the preprocessing steps taken for each task:

Atlas-to-patient brain MRI registration

The Database:IXI dataset includes 576 3D MRI scans from healthy individuals at three hospitals. In this study, we used pre-processed IXI data provided by [11]. The dataset was divided into training, validation, and test sets, with 403, 58, and 115 volumes, respectively, following a ratio of 7:1:2. Essential pre-processing procedures for structural brain MRI, such as skull stripping, resampling, and affine transformation, were performed using FreeSurfer software.[48] Subsequently, all image volumes were cropped to dimensions of 160 × 192 × 224. The alignment performance was evaluated using 29 labeled maps representing anatomical structures.

Inter-patient brain MRI registration

The OASIS dataset, consisting of 416 cross-sectional T1-weighted MRI scans, was used for inter-subject brain registration. We employed the pre-processed OASIS dataset provided by the Learn2Reg 2021 Challenge (Task 3),[49] which includes 414 three-dimensional scan images. Each MRI brain scan underwent preprocessing steps, including skull stripping, alignment, and normalization, resulting in a resolution of 160 × 192 × 224. The registration performance was evaluated using label masks of 35 anatomical structures, with the Dice Score employed as an assessment metric.

Inter-modal brain MRI registration

The BraTS 2021 dataset contains multi-modal MRI scans, including T1, T2, T2-FLAIR, and T1CE sequences, from 2000 patients. Segmentation labels are available for 1251 individuals, with a subset of 500 cases, specifically the T1 and T2 weighted images, utilized for the experiments. The dataset was partitioned into training, validation, and testing subsets, with 350, 50, and 100 images, respectively. Since the provided T1-T2 MRI images were acquired nearly simultaneously and were already aligned, we adopted the approach described in [50]. Random affine and non-affine transformations were applied to the moving and target images during training and testing. The non-affine transformation involved utilizing elastic transformations to spatially transform the moving and target images, followed by Gaussian smoothing. The degree of deformation was set at 500, with a Gaussian smoothing radius of 12. All images underwent normalization and were subsequently cropped to a size of 160 × 192 × 160.

Inter-patient liver CT registration

The LiTS dataset comprises 130 sets of training scans and 70 sets of test scans, with the labels for the test data undisclosed. In this study, we utilized the preprocessed LiTS data provided by VTN.[47] All images were cropped and zero-padded to a size of 128×128×128 around the liver, and exposure adjustments were applied to normalize the images, aligning their histograms to ensure optimal comparability.

Inter-modal abdomen MRI-CT registration

This dataset comprises 8 paired CT-MRI scans and 90 unpaired CT/MRI scans sourced from the Abdomen MRI-CT Task in the Learn2Reg challenge.[49] The scans exhibit a resolution of 192×160×192 with a voxel size of 2×2×2 mm. Each scan is accompanied by an abdominal mask, facilitating the exclusion of irrelevant regions outside the abdomen. The dataset encompasses four segmentation categories: liver, spleen, right kidney, and left kidney. For training and testing purposes, we randomly paired the unpaired CT/MRI scans to create 45 pairs, utilizing the 8 paired CT-MRI scan pairs for testing. Notably, variations in body shape and

posture can induce significant displacements between abdominal scans, necessitating that the registration model effectively accommodates large deformations.

Implementation

Our methodology was implemented using the PyTorch framework on a server equipped with an NVIDIA A100 GPU. The models underwent 500 epochs of training using the Adam optimization algorithm, with a fixed learning rate of 1×10^{-4} and a batch size of 1. No data augmentation was employed during the training process. Due to memory constraints, the MMC module comprised 4 branches with convolutional kernel sizes of 3, 5, 7, and 9. We used identical hyperparameters to ensure comparability with the work of Chen et al. (2022). Various task-specific loss functions were utilized in our experiments. Specifically, for the 3D OASIS registration (Learn2Reg task 3), we implemented the $\mathcal{L}_{sim}$, $\mathcal{L}_{reg}$, and functions. Conversely, for the remaining three tasks, only the $\mathcal{L}_{sim}$ and $\mathcal{L}_{reg}$ functions were employed. Across all tasks, the values of λ and γ were standardized to 1. The model achieving the highest Dice score on the validation set was selected for further evaluation.

Evaluation metrics

In evaluating each model's registration performance, a comprehensive examination of the overlap in anatomical and organ segmentation was conducted, with quantification using the widely utilized Dice score.[51] The Dice scores for all anatomical and organ structures were averaged across all patients, and the mean and standard deviation of these averaged scores were then compared among different registration methods. Additionally, the regularity of the deformation fields was gauged through the reporting of the percentages of non-positive values in the determinant of the Jacobian matrix of the deformation fields. The smaller the value, the better the registration ability of deformation. To further assess the performance for inter-modal brain MRI registration, inter-patient liver CT registration tasks and inter-modal Abdomen MR-CT registration, the Hausdorff distances[52] were employed to measure the degree of boundary matching between different modality labels. These meticulous evaluation metrics offer valuable insights into the performance and robustness of the proposed ScaMorph model for deformable image registration, thereby underscoring its potential in clinical and research applications.

Overall architecture

In this study, we propose the ScaMorph model for deformable image registration, as illustrated in Figure S1. The model employs a deformable image registration (DIR) network to produce a non-linear deformation field, denoted as ϕ, from the input images, which consists of the moving image $\boldsymbol{I}_m$ and the fixed image $\boldsymbol{I}_f$. The spatial transformation function[53] in the network warps the mobile image $\boldsymbol{I}_m$ to $\boldsymbol{I}_m \circ \phi$. The backbone architecture of the network is based on U-Net, integrating an encoder-decoder neural network structure with skip connections. Furthermore, the network can incorporate additional information, such as anatomical segmentation, to enhance registration accuracy. The subsequent sections comprehensively describe our neural network architecture, loss functions, and diffeomorphic variants.

Deformable registration network

The architecture of ScaMorph, depicted in Figure S2, involves the initial processing of the input data represented as $\mathcal{X} \in \mathbb{R}^{H\times W\times D\times S}$ by a reg head layer. This layer generates a sequence of 3D tokens with dimensions of $\frac{H}{H'}\times\frac{W}{W'}\times\frac{D}{D'}$. These tokens are then projected into an embedding space of dimension C. In the encoder, the size of the embedding channel C is set to 16, while H', W', and D' are all set to 2. The network consists of four stages, each with a downsampling rate of 4, 8, 16, 32. The Sca blocks in the encoder retain the same number of tokens as the input, preserving the hierarchical structure. To maintain this structure, a patch merging layer is strategically inserted at the end of each stage, reducing the resolution by a factor of 2 and resulting in a $4C$-dimensionals feature embedding. The decoder section includes consecutive upsampling and convolutional layers with a kernel size of $3\times3\times3$. During the decoding stage, each upsampled feature map is connected to the corresponding feature map in the encoding path via skip connections, followed by a decoder block. Additionally, ResBlocks[54] are strategically utilized after each output feature map of the encoder to enhance the propagation of output results from the encoding path.

Scale-aware context aggregation

Multiscale Mixed Convolution

Within the building block of our encoder, we adopt a structure similar to ViT. Instead of using the self-attention mechanism, we introduce a revolutionary Multiscale Mixed Convolution (MMC) module. MMC utilizes a convolutional modulation approach that combines the benefits of depthwise and pointwise convolutions. MMC enhances the receptive field and captures multi-scale features by dividing input channels into N heads and applying separate depthwise separable convolutions to each head (Figure S2(e)). Mathematically, MMC can be expressed as:

$$\mathrm{MMC}(\mathrm{X}) = \mathrm{Concat}\left(\mathrm{DConv}_{k_1}(x_1), \ldots, \mathrm{DConv}_{k_n}(x_n)\right), \quad \text{(Equation 2)}$$

where $\mathrm{X} = [x_1, x_2, \ldots, x_n]$ represents the input feature map, x_i indicates the i-th feature map, and k_i denotes the kernel size of the i-th feature map. Each branch uses DConv to represent depthwise convolution. To approximate conventional depthwise convolutions

with larger kernels, we employ three depthwise strip convolutions that simulate a 3D convolution with a kernel size of 5 × 5 × 5. This is achieved by applying a combination of 5 × 1 × 1, 1 × 5 × 1, and 1×1×5 convolutions in a lightweight strategy, enlarging the receptive field and enhancing the model's ability to capture long-range dependencies.

Multiscale Context Fusion

To facilitate the smooth exchange of information among multiple entities within the Multiscale Modulation and Context Fusion (MMC) framework, we introduce a lightweight module called Multiscale Context Fusion (MCF), as shown in Figure S2(e). Forming a collective by selecting one channel from each entity and utilizing the inverse bottleneck architecture, we perform an upward-downward fusion operation within each collective. This operation enhances the diversity of multiscale characteristics, and we capture the inter-group relationships using a 1×1×1 convolutional layer. The input, denoted as $G \in \mathbb{R}^{H\times W\times D\times C}$, undergoes cross-group information aggregation for all characteristics using point-wise convolution to achieve cross-fertilization of global information. The MCF module can be mathematically expressed as follows:

$$\mathrm{MCF}(G) = \mathrm{Conv}([\mathrm{Conv}(g_1), \mathrm{Conv}(g_2), \ldots, \mathrm{Conv}(g_M)]), \quad \text{(Equation 3)}$$

where $M = C/N$, and g_i represents the i-th collective. The MCF module is a streamlined version of the self-attention mechanism, effective in capturing long-range dependencies.

After capturing and aggregating multiscale spatial characteristics with MMC and MCF, we obtain an output characteristic map referred to as the modulator M. Subsequently, we utilize this modulator to modulate the value V through scalar multiplication. For the input characteristic $F \in \mathbb{R}^{H\times W\times D\times C}$, the output is computed as follows:

$$\begin{aligned} \mathrm{M} &= \mathrm{MCF}(\mathrm{MMC}(W_s F)), \\ \mathrm{Attn}(F) &= M \otimes (W_v F), \end{aligned} \quad \text{(Equation 4)}$$

where W_s and W_v represent weight matrices of linear layers, respectively. The symbol $\otimes$ denotes element-wise matrix multiplication, and *Attn* represents the attention map. Unlike self-attention, which generates an $N\times N\times N$ attention map, the modulator preserves the channel dimension. This characteristic allows spatial- and channel-specific modulation of the value after element-wise multiplication, ensuring memory efficiency, particularly when processing high-resolution images.

Loss function

The loss function for network training is derived from the energy function of conventional image registration algorithms, specifically Equation 1. This loss function consists of two main components: quantifying the similarity between the deformed moving image and the fixed image, and enforcing smoothness in the deformation field. The overall loss function for network training is defined as follows:

$$\mathcal{L}(\boldsymbol{I}_f, \boldsymbol{I}_m \circ \phi) = \mathcal{L}_{\mathrm{sim}}(\boldsymbol{I}_f, \boldsymbol{I}_m \circ \phi) + \lambda \mathcal{L}_{\mathrm{reg}}(\phi), \quad \text{(Equation 5)}$$

where $\mathcal{L}_{\mathrm{sim}}$ measures image fidelity, and $\mathcal{L}_{\mathrm{reg}}$ signifies deformation field regularization.

Measurement of image similarity

We conducted experiments using two widely adopted similarity metrics for $\mathcal{L}_{\mathrm{sim}}$. The first metric was the mean squared error (MSE), which calculates the mean of the squared differences in voxel values between $\boldsymbol{I}_f$ and $\boldsymbol{I}_m \circ \phi$:

$$\mathrm{MSE}(\boldsymbol{I}_f, \boldsymbol{I}_m \circ \phi) = \frac{1}{\Omega}\sum_{\boldsymbol{v}\in\Omega} |\boldsymbol{I}_f(\boldsymbol{v}) - [\boldsymbol{I}_m \circ \phi](\boldsymbol{v})|^2, \quad \text{(Equation 6)}$$

where $\mathbf{v}$ indicates the voxel location, and Ω represents the image domain.

Another similarity metric used was the local normalized cross-correlation between $\mathbf{I}_f$ and $\mathbf{I}_m \circ \phi$:

$$\mathrm{NCC}(\boldsymbol{I}_f, \boldsymbol{I}_m \circ \phi) = \frac{1}{\Omega}\sum_{\boldsymbol{v}\in\Omega} \frac{\left[\boldsymbol{I}_f(\boldsymbol{v}) - \mu_{\boldsymbol{I}_f}\right]\left[\boldsymbol{I}_m \circ \phi(\boldsymbol{v}) - \mu_{\boldsymbol{I}_m \circ \phi}\right]}{\sigma_{\boldsymbol{I}_f}\sigma_{\boldsymbol{I}_m \circ \phi}}, \quad \text{(Equation 7)}$$

where $\mu_{\mathbf{I}_f}$ and $\mu_{\boldsymbol{I}_m \circ \phi}$ denote the mean voxel values of $\boldsymbol{I}_f$ and $\boldsymbol{I}_m \circ \phi$ respectively, and $\sigma_{\mathbf{I}_f}$ and $\sigma_{\boldsymbol{I}_m \circ \phi}$ represent the standard deviations of $\boldsymbol{I}_f$ and $\boldsymbol{I}_m \circ \phi$ respectively.

Regularization of the Deformation Field Minimizing the L_{sim} function facilitates the approximation of $\boldsymbol{I}_m \circ \phi$ to $\boldsymbol{I}_f$, potentially resulting in a non-smooth $\circ$ operation devoid of physical realism. To enhance the smoothness of the displacement field, we apply a diffusion regularizer to the spatial gradients of the displacement $\mathbf{u}$:

$$\mathcal{L}_{\mathrm{reg}}(\phi) = \sum_{\mathbf{v}\in\Omega} \|\nabla \mathbf{u}(\mathbf{v})\|^2, \quad \text{(Equation 8)}$$

where $\mathbf{u}(\mathbf{v})$ represents the spatial gradients of the displacement field $\mathbf{u}$. We estimate spatial gradients using forward differences, specifically $\frac{\partial \mathbf{u}(\mathbf{v})}{\partial x,y,z} \approx \mathbf{u}(v_{x,y,z}+1) - \mathbf{u}(v_{x,y,z})$.

Auxiliary Segmentation Loss In addition to the loss functions for image similarity and deformation field regularization, we introduce an auxiliary segmentation loss to improve the model's performance. This loss compares the predicted segmentation mask with the

ground truth segmentation mask and is defined as:

$$\mathcal{L}_{seg}(\boldsymbol{S}_f, \boldsymbol{S}_m \circ \phi) = -\frac{1}{K}\sum_{k=1}^{K} \mathrm{Dice}\left(\boldsymbol{S}_f{}^k, \boldsymbol{S}_m{}^k \circ \phi\right) = -\frac{1}{K}\sum_k \frac{2\sum_{\boldsymbol{v}\in\Omega} \boldsymbol{S}_f{}^k(\boldsymbol{v})\left[\boldsymbol{S}_m{}^k \circ \phi\right](\boldsymbol{v})}{\sum_{\boldsymbol{v}\in\Omega}\left(\boldsymbol{S}_f{}^k(\boldsymbol{v})\right)^2 + \sum_{\boldsymbol{v}\in\Omega}\left(\left[\boldsymbol{S}_m{}^k \circ \phi\right](\boldsymbol{v})\right)^2}, \quad \text{(Equation 9)}$$

where $\boldsymbol{S}_f$ and $\boldsymbol{S}_m$ represent the organ segmentation of $\boldsymbol{I}_f$ and $\boldsymbol{I}_m$, respectively. The index k indicates the structure for $\boldsymbol{I}_f$ and $\boldsymbol{I}_m$. We employ nearest-neighbor interpolation to deform the K-channel $\boldsymbol{S}_m$ and ϕ, enabling us to compute $\boldsymbol{S}_m{}^k \circ \phi$ and propagate the gradient of $\mathcal{L}_{\text{seg}}$ back through the network. Combining $\mathcal{L}_{\text{seg}}$ with Equation 5, we obtain the objective:

$$\mathcal{L}(\boldsymbol{I}_f, \boldsymbol{I}_m \circ \phi) = \mathcal{L}_{\text{sim}}(\boldsymbol{I}_f, \boldsymbol{I}_m \circ \phi) + \lambda\mathcal{L}_{\text{reg}}(\phi) + \gamma\mathcal{L}_{seg}(\boldsymbol{S}_f, \boldsymbol{S}_m \circ \phi), \quad \text{(Equation 10)}$$

where γ is a weighting parameter controlling the strength of $\mathcal{L}_{\text{seg}}$.

Probabilistic variants

In this section, we introduce the concept of diffeomorphic registration using the ScaMorph method. The objective of diffeomorphic registration is to find a smooth and differentiable deformation field that preserves the topological characteristics of the image. To achieve this, we introduce a latent variable denoted as $\boldsymbol{z}$, which serves as a parameter for the infinitesimal motion field referred to as ϕ. The latent variable $\boldsymbol{z}$ follows a multivariate Gaussian distribution with a mean of zero and a covariance matrix denoted as $\boldsymbol{\Sigma}_z$, serving as the prior distribution $p(\boldsymbol{z})$:

$$p(\boldsymbol{z}) \sim \mathcal{N}(\boldsymbol{z}; \boldsymbol{0}, \boldsymbol{\Sigma}_z), \quad \text{(Equation 11)}$$

To achieve a diffeomorphism, as advocated in [10,23,55], we define the variable $\boldsymbol{z}$ as a stationary velocity field (SVF) responsible for governing the trajectory of the diffeomorphic deformation field $\phi^{(t)}$, where $t \in [0,1]$. The SVF undergoes an integration process over time using a scaling and squaring layer (SS) to obtain the final motion field $\phi^{(1)}$ at $t = 1$. The integration process is accomplished by recursively computing $\phi^{(1/2^t)} = \phi^{(1/2^{t+1})} \circ \phi^{(1/2^{t+1})}$, commencing from $\phi^{(1/2^T)} = \boldsymbol{p} + \boldsymbol{v}(\boldsymbol{p})/2^T$, wherein $\boldsymbol{p}$ represents a spatial location and $\boldsymbol{v}(\boldsymbol{p})$ denotes the velocity field. In our experiments, we select $T = 7$ to ensure that $\boldsymbol{v}(\boldsymbol{p})/2^T$ remains sufficiently small.

Upon obtaining the latent variable $\boldsymbol{z}$ and the motion field ϕ through the SS layer, we utilize a spatial transform layer to warp the fixed image $\boldsymbol{I}_f$ by ϕ and acquire a noisy observation of the warped image denoted as $\boldsymbol{I}_m$. This observation can be depicted as a Gaussian distribution:

$$p(\boldsymbol{I}_m|\boldsymbol{z}; \boldsymbol{I}_f) = \mathcal{N}\left(\boldsymbol{I}_m; \boldsymbol{I}_f \circ \phi, \sigma^2\mathbb{I}\right), \quad \text{(Equation 12)}$$

where σ^2 represents the variance of the additive image noise.

The main objective is to estimate the posterior distribution $p(\boldsymbol{z}|\boldsymbol{I}_m; \boldsymbol{I}_f)$ for registration to obtain the most plausible motion field ϕ for a new image pair. However, direct computation of this posterior distribution presents challenges. Consequently, we adopt a variational approach and introduce an approximate posterior distribution denoted as $q_{\boldsymbol{\psi}}(\boldsymbol{z}|\boldsymbol{I}_f; \boldsymbol{I}_m)$, which is parameterized by a fully convolutional neural network (FCN) module denoted as $\boldsymbol{\psi}$. The approximate posterior distribution is assumed to be a multivariate Gaussian distribution:

$$q_{\boldsymbol{\psi}}(\boldsymbol{z}|\boldsymbol{I}_f; \boldsymbol{I}_m) = \mathcal{N}\left(\boldsymbol{z}; \boldsymbol{\mu}_{z|\boldsymbol{I}_f,\boldsymbol{I}_m}, \boldsymbol{\Sigma}_{z|\boldsymbol{I}_f,\boldsymbol{I}_m}\right), \quad \text{(Equation 13)}$$

where the mean $\boldsymbol{\mu}_{z|\boldsymbol{I}_f,\boldsymbol{I}_m}$ and diagonal covariance $\boldsymbol{\Sigma}_{z|\boldsymbol{I}_f,\boldsymbol{I}_m}$ are learned by the FCN.

To estimate the parameters $\boldsymbol{\psi}$, our goal is to minimize the Kullback-Leibler (KL) divergence between the approximate posterior and the prior distribution, while simultaneously maximizing the evidence lower bound (ELBO)[56] of the log marginalized likelihood. This leads to the following loss function:

$$\min_{\boldsymbol{\psi}} \mathcal{KL}[q_{\boldsymbol{\psi}}(\boldsymbol{z}|\boldsymbol{I}_m; \boldsymbol{I}_f)\|p(\boldsymbol{z}|\boldsymbol{I}_m; \boldsymbol{I}_f)] = \min_{\boldsymbol{\psi}} \mathcal{KL}[q_{\boldsymbol{\psi}}(\boldsymbol{z}|\boldsymbol{I}_m; \boldsymbol{I}_f)\|p(\boldsymbol{z})] - \mathbb{E}_q[\log\, p(\boldsymbol{I}_m|\boldsymbol{z}; \boldsymbol{I}_f)] + \log\, p(\boldsymbol{I}_m|\boldsymbol{I}_f). \quad \text{(Equation 14)}$$

where $\mathcal{KL}$ denotes KL divergence, $\mathbb{E}_q$ represents the expectation concerning the approximate posterior, and $p(\boldsymbol{I}_m|\boldsymbol{I}_f)$ is the Boltzmann distribution that measures the similarity between the warped image $\boldsymbol{I}_f \circ \phi$ and the observed image $\boldsymbol{I}_m$.

The loss function comprises two terms: a reconstruction loss term and a similarity loss term that assesses the resemblance between the warped image $\boldsymbol{I}_f$ and the observed image $\boldsymbol{I}_m$. In this study, we employ a normalized local cross-correlation (NCC)[7] metric to quantify the similarity between the images. Moreover, we model $p(\boldsymbol{I}_m|\boldsymbol{z}; \boldsymbol{I}_f)$ using the Boltzmann distribution:

$$p(\boldsymbol{I}_m|\boldsymbol{z}; \boldsymbol{I}_f) \sim \exp(-\gamma NCC(\boldsymbol{I}_m, \boldsymbol{I}_f \circ \phi)), \quad \text{(Equation 15)}$$

where γ represents a negative scalar hyperparameter. Finally, the loss function can be articulated as:

$$\begin{aligned} \mathcal{L}_{kl} &= \mathcal{KL}[q_{\boldsymbol{\psi}}(\boldsymbol{z}|\boldsymbol{I}_m; \boldsymbol{I}_f)\|p(\boldsymbol{z})] - \mathbb{E}_q[\log\, p(\boldsymbol{I}_m|\boldsymbol{z}; \boldsymbol{I}_f)] \\ &= \frac{1}{2}\left[tr\left(\lambda D\boldsymbol{\Sigma}_{z|x,y} - \log\,\boldsymbol{\Sigma}_{z|x,y}\right) + \boldsymbol{\mu}_{z|x,y}^T\boldsymbol{\Lambda}_z\boldsymbol{\mu}_{z|x,y}\right] \\ &\quad + \frac{\gamma}{K}\sum_k NCC(\boldsymbol{I}_m, \boldsymbol{I}_f \circ \phi_k) + const, \end{aligned} \quad \text{(Equation 16)}$$

where $\boldsymbol{D}$ represents the graph degree matrix defined on the 3D image pixel grid, and K is the number of samples used to approximate the expectation. In our experiments, we set $K = 1$. Additionally, $\boldsymbol{\Lambda}_z$ denotes a diagonal matrix with the eigenvalues of $\boldsymbol{\Sigma}_{z|x,y}$ along the diagonal. The first term in Equation 16 corresponds to the KL divergence between the approximate posterior and the prior, while the second term represents the reconstruction loss. The constant term is independent of the parameters $\boldsymbol{\psi}$ and can be disregarded during optimization.

QUANTIFICATION AND STATISTICAL ANALYSIS

To mitigate the effects of random dataset partitioning, a five-fold cross-validation procedure was employed. Models were trained on three folds of the data, with one fold reserved for validation to identify the most effective model configuration. The remaining fold served as a test set to evaluate the model's performance. Additionally, the model's generalization ability was assessed on an external dataset. Statistical significance was established at $p < 0.05$ (two-sided) using a two-tailed test. A 95 % confidence interval (CI) was calculated for each method, incorporating Cohen's d values to quantify effect sizes.