Knowledge-Based Systems

GMmorph: Dynamic Spatial Matching Registration Model for 3D Medical Image based on Gated Mamba

--Manuscript Draft--

Manuscript Number:	KNOSYS-D-24-03551
Article Type:	Full Length Article
Keywords:	3D medical image registration, dual-branch, dynamic spatial matching, Gated Mamba module, consistency loss.
Corresponding Author:	Song Yonghong Xi'an Jiaotong University CHINA
First Author:	Lin Hao
Order of Authors:	Lin Hao
	Song Yonghong
	Li Fei
	Zhang Qi
	pe Bincheng
Abstract:	Medical image registration plays a crucial role in precise planning for radiotherapy, as well as in tasks such as image fusion and medical image analysis. Deep learning registration methods have been widely employed due to their excellent registration efficiency. However, most existing models exhibit limitations in terms of spatial matching and anatomical feature extraction between image pairs, resulting in insufficient registration accuracy and robustness towards abnormal tissues. To address this issue, we propose a dual-branch registration model architecture from the perspectives of spatial matching and feature extraction. Specifically, our approach includes: (1) Designing the Dynamic Matching module (DMM) to perform deformable and dynamic spatial matching, which generates learnable offsets based on real-time information flow in the upper and lower branches, thus promoting flexibility in searching for optimal positions; (2) Introducing the Gated Mamba Layer (GML) to achieve global feature extraction, and using channel and spatial attention to design the Detailed Enhancement Module (DEM) for enhancing local detailed features. The combination of these two modules forms the Coarse-to-Fine Feature Extraction Module (CFFEM), enabling the retrieval of coarse and fine features. Furthermore, an implicit regularization is achieved through a consistency loss, which encourages the network to balance high accuracy, low folding rate, and robustness. Our model demonstrates excellent performance in both single-modal and multi-modal image registration in semi-supervised and unsupervised modes, as validated by comparisons with various state-of-the-art methods on the IXI, OASIS, and Brats2021 datasets.
Suggested Reviewers:	Chao Fan anfan2003@163.com
	Li YaoChen yaochenli@mail.xjtu.edu.cn

Declaration of interests

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

☐The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

GMmorph: Dynamic Spatial Matching Registration Model for 3D Medical Image based on Gated Mamba

Abstract

Medical image registration plays a crucial role in precise planning for radiotherapy, as well as in tasks such as image fusion and medical image analysis. Deep learning registration methods have been widely employed due to their excellent registration efficiency. However, most existing models exhibit limitations in terms of spatial matching and anatomical feature extraction between image pairs, resulting in insufficient registration accuracy and robustness towards abnormal tissues. To address this issue, we propose a dual-branch registration model architecture from the perspectives of spatial matching and feature extraction. Specifically, our approach includes: (1) Designing the Dynamic Matching module (DMM) to perform deformable and dynamic spatial matching, which generates learnable offsets based on real-time information flow in the upper and lower branches, thus promoting flexibility in searching for optimal positions; (2) Introducing the Gated Mamba Layer (GML) to achieve global feature extraction, and using channel and spatial attention to design the Detailed Enhancement Module (DEM) for enhancing local detailed features. The combination of these two modules forms the Coarse-to-Fine Feature Extraction Module (CFFEM), enabling the retrieval of coarse and fine features. Furthermore, an implicit regularization is achieved through a consistency loss, which encourages the network to balance high accuracy, low folding rate, and robustness. Our model demonstrates excellent performance in both single-modal and multi-modal image registration in semi-supervised and unsupervised modes, as validated by comparisons with various state-of-the-art methods on the IXI, OASIS, and Brats2021 datasets.

Keywords: 3D medical image registration, dual-branch, dynamic spatial matching, Gated Mamba module, consistency loss.

1. Introduction

Constructing deformable templates for a patient cohort [1], tumor growth treatment planning, multi-atlas segmentation [2, 3]、image-guided interventions, and surgical navigation are common medical scenarios that rely on medical images from different time periods for observation. However, the variations in imaging device parameters and differences in patient posture and breathing frequency during different imaging stages make it challenging to accurately align medical images based on anatomical information. Deformable medical image registration (DIR) addresses this challenge by aligning medical images acquired at different angles and time points for further analysis and clinical processes.

Since the 1970s, research in this field has primarily focused on registering images by computing the similarity between their grayscale or features [4], Common registration methods include sequential similarity methods (including mutual information and feature point matching) [5, 6] [7] [8] [9]、multi-resolution pyramid methods [10, 11], elastic deformation methods [12, 13], Bayesian methods [14-16]. There are also diffeomorphic registration methods based on preserving topology and transform invertibility [5, 17, 18]. These methods treat registration as an iterative optimization problem by constructing metric-based objective functions. However, they are often sensitive to noise, computationally complex, and lack real-time performance [19]. Modern registration approaches based on deep learning have achieved flexibility and real-time performance. These methods directly predict deformations from a pair of images (moving and fixed) and reconstruct the registration results using spatial transformation networks (STN) [20]. Supervised learning methods [21, 22] [23-27] use ground truth deformations or image segmentation labels for training, but obtaining such supervision is time-consuming, limiting the feasibility of supervised learning methods in clinical applications [28]. Unsupervised registration methods impose similarity constraints between the registered and fixed images and utilize regularization losses to encourage networks to learn locally consistent deformation fields [29-31]. Unsupervised methods have become the main research direction in the field due to their simplicity and lack of label requirements. For example, Zhao et al. [32] designed an end-to-end recursive cascade network to iteratively improve registration results, decomposing large deformations into multiple small deformations. Kim et al. [31] proposed a multi-scale deformable image registration method by introducing cycle consistency. Ma et al.[33] proposed a non-iterative coarse-to-fine registration network (NICE-Net) to avoid the complexity of cascaded or recursive networks. These works, based on CNNs, generate deformable fields from coarse to fine using different multi-scale registration strategies. However, the locality of CNNs limits their ability to capture large displacement deformations and lacks a spatial matching process. Subsequently, researchers have utilized Transformers to capture long-range positional dependencies between image patches [34] [29, 35]. In particular, some works have added Cross-attention (CA) on top of Self-attention (SA) in VIT-based registration methods to perform spatial matching. For example, Shi et al. proposed Xmorpher, a dual-parallel feature registration network, by constraining the attention calculation between different-sized base windows and search windows [36]. Chen Chen et al. extract features from each image branch and explicitly match the registration image features using Transformers [37]. However, the CA mechanism is rigid in terms of spatial matching: either computing CA globally [25], which hinders hierarchical feature extraction and is only suitable for low-resolution features, or computing CA within a fixed but expanded window [23], which significantly increases computational complexity. Moreover, the quadratic complexity of VIT prevents it from working in shallow layers of the network, making it unable to establish accurate spatial dependencies and capture pixel-level anatomical information. Finally, there are also works based on GANs [38, 39], automated learning, and meta-learning registration frameworks [40] [41]. These methods innovate in terms of registration patterns but often face problems such as complex hyperparameters and loss designs, unstable working modes, and appearance uncertainty caused by GANs.

There are several challenges in registration tasks: （(1) How to alleviate the issue of abnormal pixel transformation in the deformation field (deformation folding problem). In complex deformation scenarios with rich details, such as whole-brain images, abundant displacements can easily lead to folding problems. Regularization losses partially alleviate this issue, but anatomical uncertainties still exist in the registered images. Scholars have combined deep learning methods with traditional diffeomorphic algorithms [42] to ensure a one-to-one mapping between the topology and the source and target images. The diffeomorphic algorithm's continuous reversibility completely avoids folding phenomena but at the cost of reduced accuracy. (2) Medical images often contain tissue information from other body parts (referred to as abnormal tissues) due to the patient's non-standard positioning during the acquisition process. This information can mislead the registration network and result in unreasonable deformations. Therefore, the model should possess a certain robustness when encountering such scenarios.

(1) Recently, the state-space model (SSM) [43] has attracted considerable attention among researchers. Building upon SSM, Mamba [44] establishes not only long-range dependencies but also linear complexity in terms of input size, making it suitable for registration tasks. (2) We believe that spatial information matching is more crucial than image semantics in registration tasks; thus, excessively deep network designs are unnecessary. (3) Consistency constraints can achieve implicit regularization and promote topological preservation. Inspired by these insights, we propose a dual-path dynamic spatial matching registration model called Gmmorph, based on the gated Mamba. The model has a symmetric dual-branch structure with five layers, enabling interactive registration between the fixed and moving images. The design of the consistency loss constrains and promotes the two registration processes, alleviating the folding phenomenon in the deformation field. Additionally, the Dynamic Matching Module (DMM) is designed based on the dimension of spatial matching, and the Coarse-Fine Feature Extract Module (CFFEM) is designed based on Mamba, which includes the Gated Mamba Layer (GML) and Detail Enhance Module (DEM).

Specifically, our contributions can be summarized as follows:

1. We propose a symmetric dual-path dynamic spatial matching registration framework. The interactive registration mode allows easy application of consistency constraint loss to achieve reversibility constraint on the deformation field, thereby improving the folding phenomenon. The design of the model structure and loss enables it to possess stronger robustness when dealing with abnormal tissue information. We also design an evaluation factor, D-S, which provides more accurate registration results for datasets containing abnormal tissue data.
2. We design the Dynamic Matching Module (DMM) to perform spatial matching at each layer between the branches. The module adopts a deformable mechanism with learnable offsets. This allows the sampling range of the reference image to take any shape based on the offset, enabling the model to find the optimal solution beyond the sampling range without consuming additional computational resources.
3. We design the Gated Mamba Layer (GML) in the gated Mamba module to achieve global feature extraction. The gating mechanism has high flexibility in handling different types of sequential data. It can adaptively learn how to retain or forget information based on the characteristics of the data. The linear complexity of Mamba enables global modeling of pixel-level features. The DEM module is designed with channel and spatial attention to enhance local detailed features. The combination of the two forms the CFFEM module, which achieves feature querying at different scales.
4. The model is applicable to both semi-supervised and unsupervised modes. In the semi-supervised mode, which

targets datasets with segmentation labels, the design of the label-crossed input data further improves the performance of the semi-supervised mode. The model is compared with various state-of-the-art methods on the IXI, OASIS, and Brats2021 datasets, and the results demonstrate its excellent performance in single-modal and multi-modal image registration in various registration scenario.

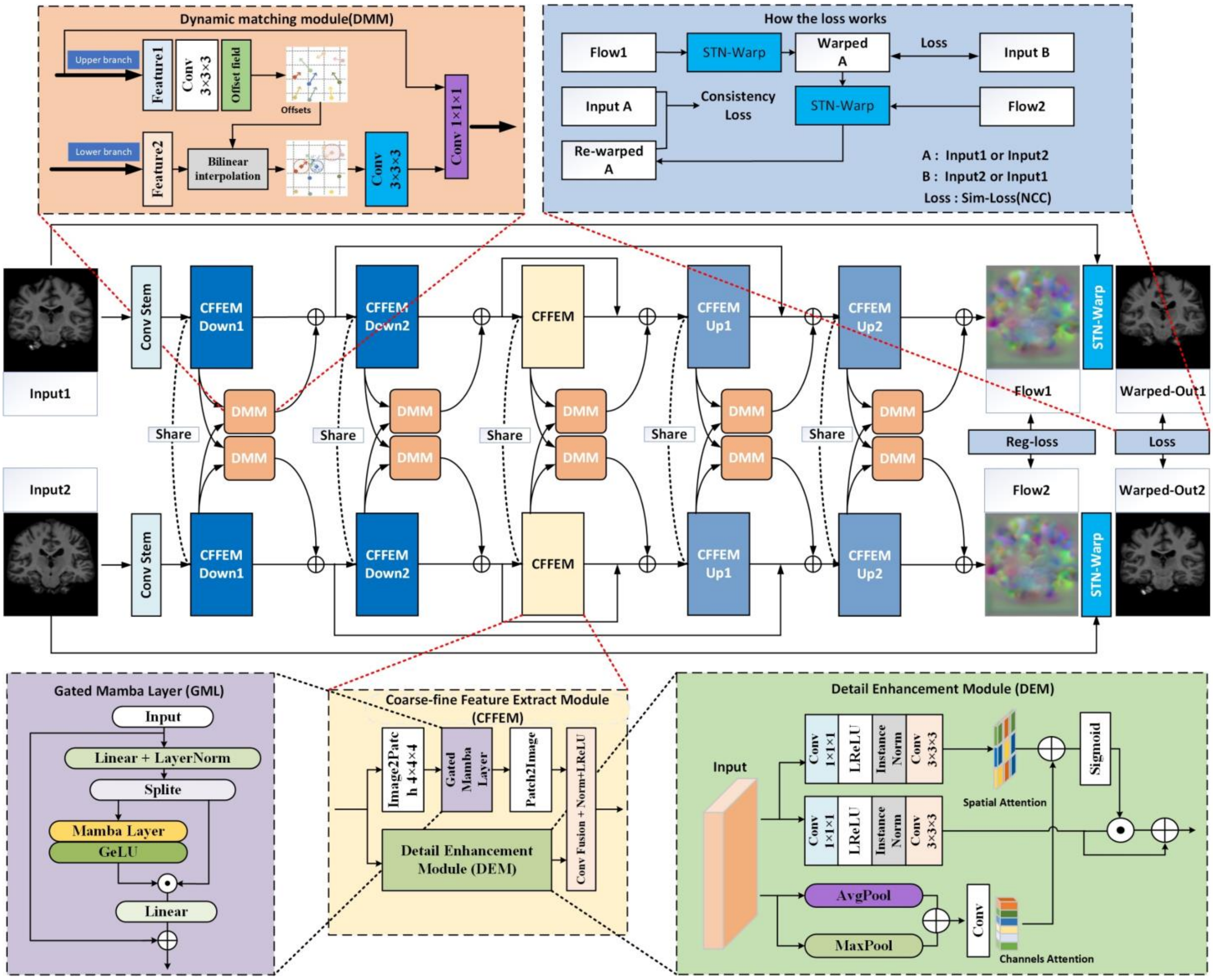


Fig. 1: Overall framework of BDMS. Flow represents the deformation field. STN-Warp represents using the spatial transformation network to deform images. The upper branch registers with Input 2 as the fixed image, while the lower branch registers with Input1 as the fixed image. Information interaction between branches is achieved through DDM. ”share” indicates parameter sharing.

2. Methond

Deep learning-based image registration methods are typically implemented within a variational framework that addresses energy minimization in the deformation space. Let F, M be fixed and moving volumes defined over a 3-D mutual spatial domain $\Omega \subseteq R^3$. The energy equation for the registration problem can be expressed as:

$$\hat{\phi} = \arg\min_{\phi} E(F, M \circ \phi) + \lambda R(\phi) \tag{1}$$

ϕ represents the deformation field that maps M nonlinearly to F. $M \circ \phi$ utilizes a spatial transformation network (STN) to apply the deformation field to M, generating the registered image. R typically represents a regularization function for ϕ (often L1 or L2 norm regularization) to prevent folding artifacts in the deformation field, while E represents a similarity loss between F and M. The essence of the registration problem lies in aligning F and $M \circ \phi$ as closely as possible within the anatomical region while ensuring that the learned ϕ possesses a smooth topology.

The architecture of Gmmorph is depicted in Figure 1. The model adopts a dual-branch interactive registration structure, inspired by interactive registration that facilitates multi-perspective observations for practical diagnosis and enables the design of consistency loss between different directions of ϕ. The two downsampling operations and the five-layer design concentrate features at the pixel level while retaining the ability to handle large displacements. The DMM facilitates information transfer between the upper and lower branches, learning a deformable mechanism that involves a certain offset from the current branch and dynamically selects and matches spatial information from the opposing branch. GML enhances the flexibility of global information extraction through a gating mechanism, while the DEM supplements anatomical details under the guidance of global information using channel and spatial attention. CFFEM handles the task of feature extraction. The consistency loss connects the upper and lower branches, minimizing the generation of uncertain pixels.

2.1 Dynamic matching Module

The upper and lower branches dynamically match and query each other based on their "own conditions," forming channels and modes of information interaction. This process is realized in the DMM. Taking the upper branch DMM as an example, specifically, the features of the upper branch undergo a 3x3x3 convolution to generate the offset vector Δc, given by:

$$\Delta c(p_0) = \sum_{p_o \in R} w(3 \times 3 \times 3) \cdot x(p_0(\mathrm{c})) \qquad (2)$$

p_0 represents the feature point corresponding to the feature map x of the upper branch, c represents the position coordinates of p_0, and $w(3 \times 3 \times 3)$denotes the convolution kernel. In a single convolution operation, Δc has a scale of 3x3x3, indicating the displacement along the x, y, and z dimensions of p_0. Subsequently, using trilinear interpolation, the displaced position c+Δc is used to obtain the corresponding feature of the lower branch, given by:

$$y(q_0(c + \Delta c)) = \sum_{q_0(c)} G(q_0(c), q_0(c + \Delta c)) \cdot y(q_0(\mathrm{c})) \qquad (3)$$

$$G(q,p) = g(q_x, p_x) \cdot g(q_y, p_y) \cdot g(q_z, p_z) \qquad (4)$$

$q_0(c)$ represents the feature point corresponding to the feature map y of the lower branch at position coordinates c. G is a 3D bilinear interpolation kernel, and $g(a,b) = max(0,1 - |a - b|)$. The use of 3D bilinear interpolation is to address the pixel value of the corresponding pixel when c+Δc is a decimal. Finally, the output of DMM is given by:

$$DMM_{output} = \sum_{p_0 \in R} w(1 \times 1 \times 1) \cdot y(q_0(c + \Delta c)) \qquad (5)$$

The DMM_{output} and the upper branch features x undergo spatial matching and interactive fusion through a 1x1x1 convolution layer. The entire process, represented by Figure 2 in practical module operation, enables the DMM to "understand" the feature distribution of the upper and lower branches, enhancing the flexibility of information extraction.

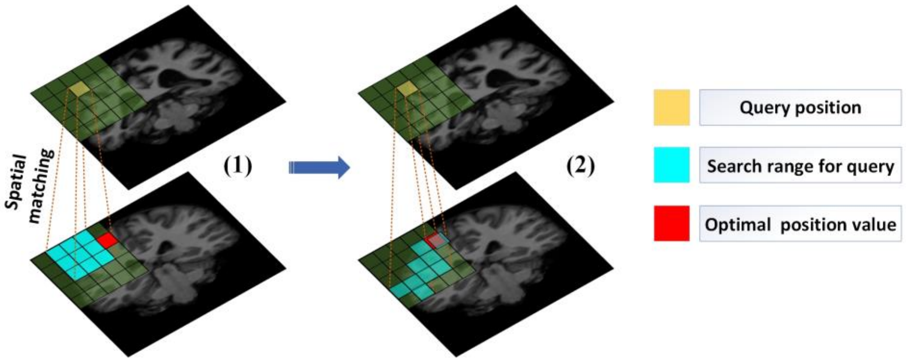


Fig. 2: With the DDM module, spatial matching changes from ”hard matching” at the corresponding position to ”soft matching” where the search region follows the feature changes. (1) represents that without the DDM module, the best corresponding position (red square) for the yellow square is not in the blue region, indicating that the optimal solution is not in the search space. (2) illustrates that the DDM dynamically deforms the search region based on the generated offset, thus finding the best corresponding position (red square).

2.2 Gated Mamba Layer

Due to the quadratic complexity of Transformers, their ability to process high-resolution images is indirectly reduced, making them unsuitable for shallow layers in a network. However, shallow features are more conducive to expressing spatial information of pixels, which is beneficial for registration tasks. Recent works, such as Vim [45] and VMamba[46], have used SSM to achieve linear complexity and global receptive fields, and have completed tasks such as image classification and segmentation on natural images. Naive SSM [47] can map a one-dimensional function or sequence u(t) to y(t)∈R and can be represented by the following linear ordinary differential equation (ODE):

$$x'(t) = \mathbf{A}x(t) + \mathbf{B}u(t) \qquad (6)$$

$$y(t) = \mathbf{C}x(t) \tag{7}$$

where state matrix $A \in R^{N \times N}$ and B, $C \in R^{N}$ are its parameters and $x(t) \in R^{N}$ denotes the implicit latent state. Naive SSM can be intuitively seen as an input-dependent variant system of RNN, providing linear complexity, but it is more difficult to train.

The Structured State Space Sequence Model (S4) [48] significantly improves the naive SSM by imposing structural forms on the state matrix A and introducing efficient algorithms. Specifically, the state matrix is designed and initialized using the High-order Polynomial Projection Operator (HIPPO) [49], enabling the construction of deep sequence models [50] with rich capacity and efficient long-range reasoning ability. Mamba changes the discrete modeling approach of SSM and adopts a different input-dependent selection mechanism from traditional time and input-invariant SSM. The SSM parameters are parameterized, allowing for flexible and efficient information selection in the input. In addition, combining with the scan cycle computation model, Mamba can perform fast inference on GPUs by linearly extending with the length of the sequence.

We propose the GML to enhance the original Mamba layer for deep spatial feature extraction in images. Given an input with a shape of $B \times C \times D \times H \times W$, it is divided into a sequence of image blocks with a size of 4^3 and position embeddings through Image2Patch in the CFFEM module. After passing through the Liner layer in GML, project tokens with a shape of $B \times (D/4 \times H/4 \times W/4) \times 2C$ are obtained. After channel-wise segmentation, they go through the Mamba and GeLU non-linear activation layers. With the introduced gating mechanism, it can adaptively learn how to retain or forget information. Finally, after the Patch2Image operation, it returns to the input resolution.

2.3 Detail Enhance Module

To complement anatomical feature details, we introduce a Channel and Spatial Attention Dual Encoding Module (DEM). Given an input x, channel attention can be represented as:

$$CA = Conv(AvgPool(x) + MaxPool(x)) \tag{8}$$

MaxPool is beneficial for extracting salient features from the feature maps, while AvgPool can model a wider range of information. For spatial attention, we obtain the Spatial Map by compressing the channel dimension to enhance the spatial correlation of anatomical information. It can be represented as:

$$SA = Conv(Norm(LReLU(Conz(x)))) \tag{9}$$

The channel and spatial attention are applied to the incoming feature information and introduce skip connections to prevent information loss. Finally, the output features of DEM and GML undergo coarse-to-fine fusion in the last layer of the Cross-Field Feature Embedding Module (CFFEM).

2.4 Loss Functions

The loss function consists of several components, including the similarity constraint $\mathcal{L}_{\text{sim}}$ between the registered image and the fixed image, the consistency loss $\mathcal{L}_{\text{con}}$ based on $\mathcal{L}_{\text{sim}}$, and the regularization term $\mathcal{L}_{\text{reg}}$ to regularize the deformation field. The process of the loss function is illustrated in Figure 1, where the inputs are denoted as A and B. The similarity constraint:

$$\mathcal{L}_{\text{sim}}(A, B, \phi^{A \to B}, \phi^{B \to A}) = 0.5\,\text{NCC}\big(A, (B \circ \phi^{B \to A})\big) + 0.5\text{NCC}\big(B, (A \circ \phi^{A \to B})\big) \tag{10}$$

where NCC denotes the normalized cross-correlation loss:

$$\text{NCC}\big(B, (\text{A} \circ \phi^{A \to B})\big) = \sum_{\mathbf{p} \in \Omega} \frac{\left(\sum_{\mathbf{v}_i} (\text{B}(\mathbf{p}_i) - \bar{B}(\mathbf{p}))([\text{A} \circ \phi](\mathbf{p}_i) - [\bar{A} \circ \phi](\mathbf{p}))\right)^2}{\left(\sum_{\mathbf{p}_i} (B(\mathbf{p}_i) - \bar{B}(\mathbf{p}))^2\right)\left(\sum_{\mathbf{p}_i} ([A \circ \phi](\mathbf{p}_i) - [\bar{A} \circ \phi](\mathbf{p}))^2\right)} \tag{11}$$

where $\bar{A}(\mathbf{p})$ and $\bar{B}(\mathbf{p})$denotes themeanvoxel valuewithin the local window of size n^3 centeredat voxel **p**. We used n = 9 in the experiments. $\bar{A} \circ \phi$ denotes the deformation of $\bar{A}$ using $\bar{B}$as the fixed image.

The consistency loss:

$$\mathcal{L}_{\text{con}}\left(A, B, A_{seg}, B_{seg},, \phi^{A \to B}, \phi^{B \to A}\right) = 0.5NCC\big(B, (B \circ \phi^{B \to A} \circ \phi^{A \to B})\big) + 0.5\,NCC\big(A, (A \circ \phi^{A \to B} \circ \phi^{B \to A})\big)$$

This loss ensures the implicit reversibility of the registration process, promoting the reversibility and smoothness of the ϕ field. Its implicit flexibility allows the model to handle pixel-wise non-local jumps caused by preserving registration details, i.e., folding phenomenon. Furthermore, this loss penalizes factors that affect the differentiability of the forward and backward deformation fields, such as noise and anomalous anatomy.

In this work, NCC is only used for single-modal image registration. For multi-modal image registration, measures such as MIND[51] or NMI [52] are employed, which capture the differences in self-similarity or mutual information between images.

The regularization term $\mathcal{L}_{\text{reg}}$ prevents local inconsistencies in the ϕ field and is defined as:

$$\mathcal{L}_{\text{reg}}\,(\phi^{A\rightarrow B},\phi^{B\rightarrow A}) = \sum_{P\in\Omega} 0.5(\,||\nabla\phi^{A\rightarrow B}(P)||^2 + ||\nabla\phi^{B\rightarrow A}(P)||^2) \tag{12}$$

This loss globally enforces the smoothness of the deformation field, ensuring that the registration process progresses in the correct direction.

2.5 Semi-supervised registration mode

When dealing with data that includes segmentation labels, the semi-supervised constraint of the labels serves as the optimal parameter update direction for the registration model. Unlike the additional segmentation network mentioned in VoxelMorph, which uses the registration labels and ground truth for loss, we directly utilize the dice loss of the segmentation labels to constrain the registration network. Additionally, we employ an "interleaved input" mode for both the images and labels, as shown in Figure 3, with the aim of "softly constraining" the pixel displacement range using the anatomical regions annotated by the labels. As shown in Figure 4, during the registration process, the white tissue region aligns to the yellow region, and the "interleaved input" allows the network to "know" that the pixel movement range for the current branch corresponds to the yellow region. Furthermore, the semi-supervised mode introduces the $\mathcal{L}_{\text{Dice}}$ loss:

$$\mathcal{L}_{\text{Dice}}\left(A_{seg},B_{seg},\phi^{A\rightarrow B},\phi^{B\rightarrow A}\right) = 0.5\text{Dice}\big(A_{seg},(B_{seg}\circ\phi^{B\rightarrow A})\big)+0.5\text{Dice}\big(B_{seg},(A_{seg}\circ\phi^{A\rightarrow B})\big) \tag{13}$$

Where:

$$\text{Dice}\big(B_{seg},(A_{seg}\circ\phi^{A\rightarrow B})\big) = 1-\frac{1}{K}\sum_k \frac{2\sum_{\mathbf{p}\in\Omega} B_{seg}{}^{k}(\mathbf{p})[A_{seg}{}^{k}\circ\phi](\mathbf{p})}{\sum_{\mathbf{p}\in\Omega}\left(B_{seg}(\mathbf{p})\right)^2+\sum_{\mathbf{p}\in\Omega}\left([A_{seg}{}^{k}\circ\phi](\mathbf{p})\right)^2} \tag{14}$$

$A_{seg}{}^{k}$ represents the one–hot representation of the multi–class label A_{seg}, where k denotes the number of classes, i.e., the original label is converted into a binary label with K channels. It is worth noting that $\mathcal{L}_{\text{Dice}}$ has a promoting effect on the activation of "soft constraints".

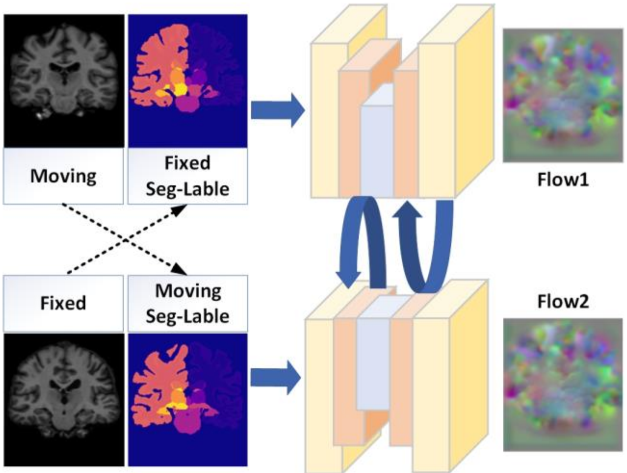


Fig. 3: "Cross input" mode for images and labels

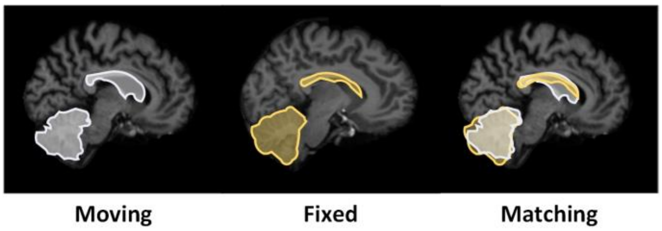


Fig. 4 Soft constrainting process, where the white and yellow regions represent the same brain tissue. The white region should align with the yellow region, meaning the movement range of pixels in the white region corresponds to the white region

3. Experiment

3.1 Datasets

OASIS: We obtained 375 3D-MRI-T2 images from the 2021 Learn2Reg Challenge (Hering et al. [53]) for the patient-to-patient task. Among them, 355 and 20 images were used for the training and testing sets, respectively, while the validation set remained the same as the testing set. The standard preprocessing protocol for brain structural MRI, including skull stripping, resampling, and affine transformation, was performed using Fischl (2012) [54]. All images were cropped to a size of 160×192×224 and normalized between 0 and 1. For each training input of the model, we randomly selected two patient scans from the training set. During training, we selected 35 significant brain regions and the background region for computing the Dice loss and the DSC (Dice similarity coefficient) metric on the validation set. During testing, the DSC metric was not computed for the background region.

IXI: We selected 552 T1-weighted 3D brain MRI images from the Information eXtraction from Images (IXI) database for the atlas-to-patient task. The dataset was divided into training, validation, and testing sets in an 8:1:1 ratio. A random image volume from the dataset was chosen as the fixed image, while the Atlas [31] image volume served as the moving image. The dataset was preprocessed following the same protocol as OASIS. The dataset provided annotations for 45 brain regions, with 40 regions labeled in the Atlas template image. During training, we computed the Dice loss and DSC metric for the 40 brain regions and the background region based on the annotations in the Atlas. During testing, we selected 30 significant brain regions for computing the DSC metric, which aligns with the testing scheme of most registration methods to avoid inaccuracies caused by mislabeling or missed labeling of non-significant regions.

Brats2021: The BraTS21 dataset refers to the 2021 Multimodal Brain Tumor Segmentation Challenge (BraTS21) dataset, which consists of 1251 brain MRI scans with four different modalities (T1, T1-CE, T2, FLAIR). The four modalities of each patient's scans are already aligned, and skull removal has been performed on each scan. The resolution of the images is 1mm×1mm×1mm. We randomly selected 300 scans for the training set and 10 pairs for the testing set in each modality. The selected images were cropped to a size of 160×160×128. Only the tumors are annotated in this dataset. Different modalities of scans from different patients were used for multimodal registration. Due to the significant variations in tumor size and location among different patients, the registration accuracy is not suitable to be measured using the DSC of tumor segmentation labels.

3.2 Evaluation Metrics

DSC(Dice Similarity Coefficient) : The DSC quantifies the overlap between different tissue regions in the segmentation labels of an image. A high DSC indicates a substantial alignment of the registered image with the fixed image's tissue regions.

$$\mathrm{Dice}\left(S_{M(\Phi)}^{K}, S_{F}^{K}\right) = 2 \cdot \frac{\|S_{R(\Phi)}^{K} \cap S_{F}^{K}\|}{\|S_{R(\Phi)}^{K}\| + \|S_{F}^{K}\|} \tag{15}$$

Where S_F^K and S_R^K are seg-labels of fixed image and warped image on organ K respectively.

SSIM: The SSIM [55] between the registered image and the fixed image reflects the consistency in terms of texture, details, and structural range. A high SSIM suggests that the registration model has effectively deformed the images.Given warped image R and fixed image F:

$$\mathrm{SSIM} = L(R,F) \times C(R,F) \times S(R,F) = \frac{(2\mu_R\mu_F+\xi_1)(2\sigma_{RF}+\xi_2)}{(\mu_R^2+\mu_F^2+\xi_1)(\sigma_R^2+\sigma_F^2+\xi_2)} \tag{16}$$

L, C, R represent luminance, contrast, and structural similarity, respectively, while μ represents the mean and σ represents the variance or covariance of the images.

HD95：95th Hausdorff Distance[56] provides a global assessment of the differences in the registration results. It reflects the maximum difference between the registered image and the fixed image's structures and is effective in detecting outliers in the registered image. A smaller HD95 indicates less local displacement or boundary shape differences.

$$H(\mathrm{S}_R, \mathrm{S}_F) = max(h(\mathrm{S}_F, \mathrm{S}_R), h(\mathrm{S}_M, \mathrm{S}_F)) \tag{17}$$

$$h(\mathrm{S}_R, \mathrm{S}_F) = \max_{a \in \mathrm{S}_M} \left\{ \min_{b \in \mathrm{S}_F} \| \mathrm{s}_r - \mathrm{s}_f \| \right\} \tag{18}$$

Where S_F and S_R are seg-labels of fixed image and warped image

Jacobian Matrix： The Jacobian matrix $J_\phi(\mathbf{p}) = \nabla\phi(\mathbf{p}) \in R^{3\times3}$ （p=(i, j, k)） measures the rate of change of the deformation field at a given point, capturing voxel-level deformations. Volume contraction: $J_\phi(p) \in [0,1]$; volume expansion: $J_\phi(p) > 1$. A non-positive determinant of the Jacobian matrix indicates a locally non-invertible transformation, i.e., a folding point. We compare the registration model's anti-folding performance by counting the number of voxels where $J_\phi(\mathbf{v}) \leq 0$.

$$
|J_\phi(i,j,k)| = \left\| \begin{matrix} \frac{\partial i}{\partial x} & \frac{\partial j}{\partial x} & \frac{\partial k}{\partial x} \\ \frac{\partial i}{\partial y} & \frac{\partial j}{\partial y} & \frac{\partial k}{\partial y} \\ \frac{\partial i}{\partial z} & \frac{\partial j}{\partial z} & \frac{\partial k}{\partial z} \end{matrix} \right\| \tag{19}
$$

NMI: Normalized Mutual Information measures the similarity between two topologies from the perspective of mutual information. In the evaluation of multimodal registration, it reflects the overall alignment between the warped image R and the fixed image F.

MAE: The Mean Absolute Error reflects the average absolute pixel distance between the warped image R and the fixed image F. A smaller MAE indicates a closer alignment between R and F.

$$
\text{MAE} = \frac{\sum_{i \in R} |R_i - F_i|}{\text{n}} \tag{20}
$$

n represents the total number of pixels, and R_i represents the grayscale value of R at pixel i.

D-S: When the registered image is faced with registration environments that include abnormal tissues (e.g., the yellow circle in Figure 5), a high SSIM may not reflect the accuracy of the registration. To comprehensively evaluate the registration accuracy and the deformability of the network, we combine DSC and SSIM to design the D-S metric.

$$
D - S = DSC \cdot \left(SSIM^{\alpha \cdot (1 - DSC) \cdot DSC}\right) \tag{21}
$$

α is a tuning factor that controls the degree of inclination of the D-S value towards DSC. In this study, it is set to 2. Through this formula, the contribution of the SSIM value to the final comprehensive score is determined based on the DSC value. When DSC is high, SSIM contributes more to the growth rate of the comprehensive score. When DSC is low, SSIM contributes less to the growth rate of the comprehensive score.

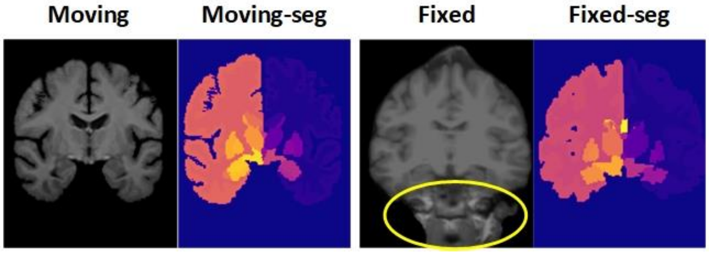


Fig. 5 Slices in the z-axis direction of the IXI dataset, where ”seg” represents the semantic segmentation masks of the images. Different colors indicate different organs (as annotated in the regions). The yellow ellipses denote regions of abnormal tissue. The parameter updates of the registration network are driven by the differences between the warped image and the fixed image. Specifically, a similarity loss is established between the warped and the fixed image. Therefore, these abnormal tissues can impact the network’s ability to produce correct displacements.

In addition, we also provided the standard deviations (STDs) of each metric in the experimental results to evaluate the stability of these registration models.

3.3 Baseline methods and Implementation Details

When evaluating the performance of our proposed model, we compared it with the classical registration model VoxelMorph [42] , as well as recently proposed registration models with publicly available code that have had significant impact in the past two years, namely TransMorph [29] , TransMatch [57] , and Xmorpher [36]. VoxelMorph is currently the most widely used baseline registration model, employing a U-net architecture that is simple yet effective and widely recognized. TransMorph, introduced in 2022, explores the application of VIT in registration models. TransMatch and Xmorpher are recent works that incorporate spatial matching mechanisms or cross-attention mechanisms to improve registration accuracy. These works are briefly introduced in the introduction, and their original papers have demonstrated the advantages of deep learning-based methods over traditional registration algorithms. Therefore, we did not select traditional registration algorithms for comparison. When reproducing each model, we kept the hyperparameters and loss settings consistent with the original papers and trained each model for 300 epochs on each dataset, saving the model with the best validation accuracy.

The comparison algorithms and our proposed model, GMmorph, were trained and tested on an NVIDIA RTX A800 GPU, using PyTorch 2.0.1 on the Ubuntu operating system. We used the Adam optimizer with a learning rate set to 0.0002. Our proposed GMmorph model converged very quickly, and we only trained it for 60 epochs on each dataset.

3.3 Comparative experimental results

We conducted single-modal registration tasks on the OASIS and IXI datasets. The visual analysis of the inter-patient task on the OASIS dataset is shown in Figure 6, and the mean and standard deviation of each metric are presented in Table 1. The visual analysis of the atlas-to-patient task on the IXI dataset is shown in Figure 7, and the objective analysis is summarized in Table 2. To assess the registration accuracy of specific brain organs, we present the registration accuracy (measured by the DSC metric) of randomly selected 10 brain tissues in Figure 8.

From Figure 6, it can be observed that GMmorph exhibits more similarities to the details indicated by the yellow markers in the Fixed image. The results in the Warped-seg column demonstrate that our method can accurately align small tissue regions, with the registered label shapes being more similar to those of the fixed image labels. By carefully observing the Flow column, it can be discerned that the deformation field itself exhibits more localized displacements. In the Jacobian column, GMmorph shows minimal folding regions while maintaining high-intensity deformations. The difference images after registration provide a more intuitive representation of the similarity between the warped result and the fixed image, with the Grid-flow demonstrating smoother displacements that are not adhered together in different directions. From Figure 7, it can be seen that most of the selected comparison methods suffer from severe folding artifacts and produce unwanted tissues in the registered images when faced with complex registration scenarios involving abnormal organs. In comparison, Transmatch and our method perform relatively well. However, from the Grid-flow and Jacobian columns, it can be inferred that GMmorph exhibits more diverse displacements and maintains a lower folding region. The other three comparison methods all exhibit adhered displacements in the Grid-flow column.

In terms of objective evaluation, as observed in Table 1, the performance of each metric is optimal, with the model achieving high SSIM while aligning relevant anatomical regions as much as possible. It is worth noting that GMmorph only has 10,420 folding points out of approximately 6.88 million (160×192×224) voxels. From Table 2, it can be seen that GMmorph performs comparably to Transmatch in terms of HD95, while VoxelMorph achieves a suboptimal SSIM but with a significantly larger number of folding points. Overall, the objective evaluation results indicate that the model has excellent deformation capabilities, achieving high registration accuracy while keeping the number of folding points low.

The box plots in Figure 8 demonstrate that GMmorph achieves optimal registration accuracy for individual brain tissues. From Figure 9(a)(b), it can be observed that GMmorph exhibits extremely fast convergence and a smoother training curve, indicating the stability of the model itself. According to the parameter comparison in Figure 9(c), GMmorph achieves satisfactory parameter efficiency, with a much lower parameter count than Transmatch, which has a slightly lower registration accuracy, and comparable to Xmorpher but with significantly higher accuracy.

Table 1: Objective evaluation of each model on OASIS (red is optimal, blue is sub-optimal)

OASIS	**Comprehensive**			**Haus Distance**	**Jacobian<0**
Affine DSC: 0.572	**DSC**	**SSIM**	**D-S**		
VoxeMorph	0.784±0.026	0.9565±0.004	0.772±0.020	2.3071±0.5471	61091±11773
TransMorph	0.770±0.026	0.9437±0.004	0.754±0.020	2.3472±0.5727	62349±10895
TransMatch	0.802±0.022	0.9593±0.0041	0.791±0.017	2.0269±0.5579	53015±8646
XMorpher	0.778±0.019	0.9169±0.005	0.755±0.016	2.2368±0.4921	57508±7677
GMmorph	0.829±0.022	0.9710±0.0036	0.822±0.017	1.8512±0.4636	10420±3695

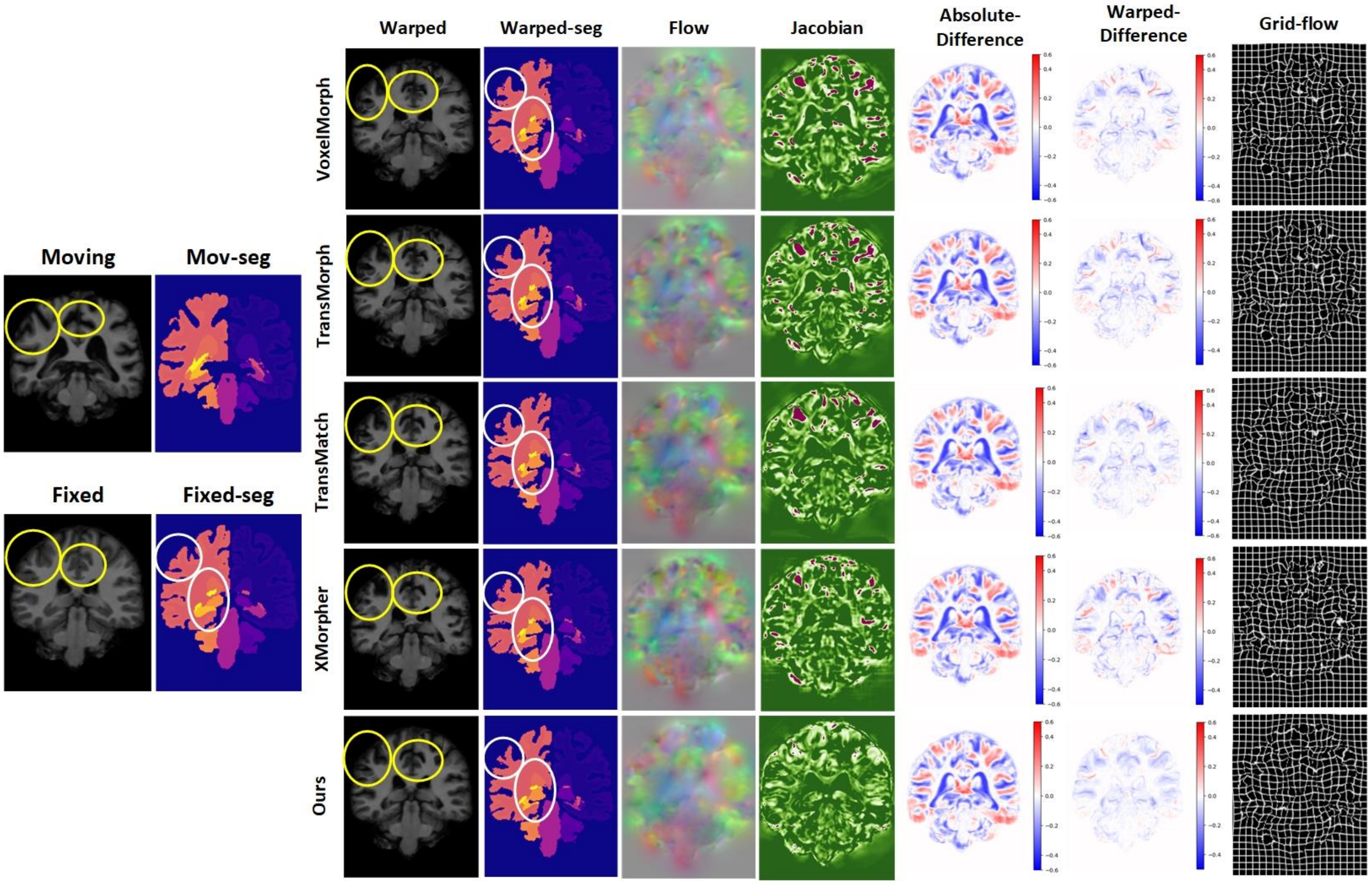


Fig. 6 The circled areas are of particular interest, where we aim for "Warped" and "Warped-seg" to align as closely as possible with "Fixed" and "Fixed-seg". In the "Flow" (deformation field) column, we visualize the displacement in the x, y, and z directions of the deformation field as RGB color channel intensities. Different colors represent different displacement directions, and the clarity of the "Flow" contours indicates whether the deformation field is excessively smooth. The "Jacobian" column visually represents the Jacobian determinant obtained from the deformation field, where red indicates folding points, white represents deformed regions, and green represents unchanged regions. "Absolute Difference" and "Warped Difference" depict the difference images between the Moving image before and after registration with the fixed image. Closer to white indicates smaller differences, while red and blue represent positive and negative differences, respectively. The "Grid-flow" column visualizes the deformed grid image, providing an intuitive observation of the local displacement direction of the deformation field. Adhesion between grids indicates folding points.

Table 2: Objective evaluation of each model on IXI (red is optimal, blue is sub-optimal)

IXI	**Comprehensive**			**Haus Distance**	**Jacobian<0**
Affine DSC: 0.413	**DSC**	**SSIM**	**D-S**		
VoxeMorph	0.734±0.031	0.9278±0.0144	0.713±0.024	5.9856±0.9943	114636±29452
TransMorph	0.731±0.030	0.9160±0.0182	0.706±0.024	5.9558±0.8536	100601±22920
TransMatch	0.744±0.029	0.9036±0.0182	0.716±0.023	5.1902±0.8100	28528±4741
XMorpher	0.714±0.033	0.8993±0.0191	0.684±0.026	5.7779±0.8354	118072±21941
GMmorph	0.764±0.030	0.9323±0.0189	0.745±0.024	5.269±0.8415	26637±8002

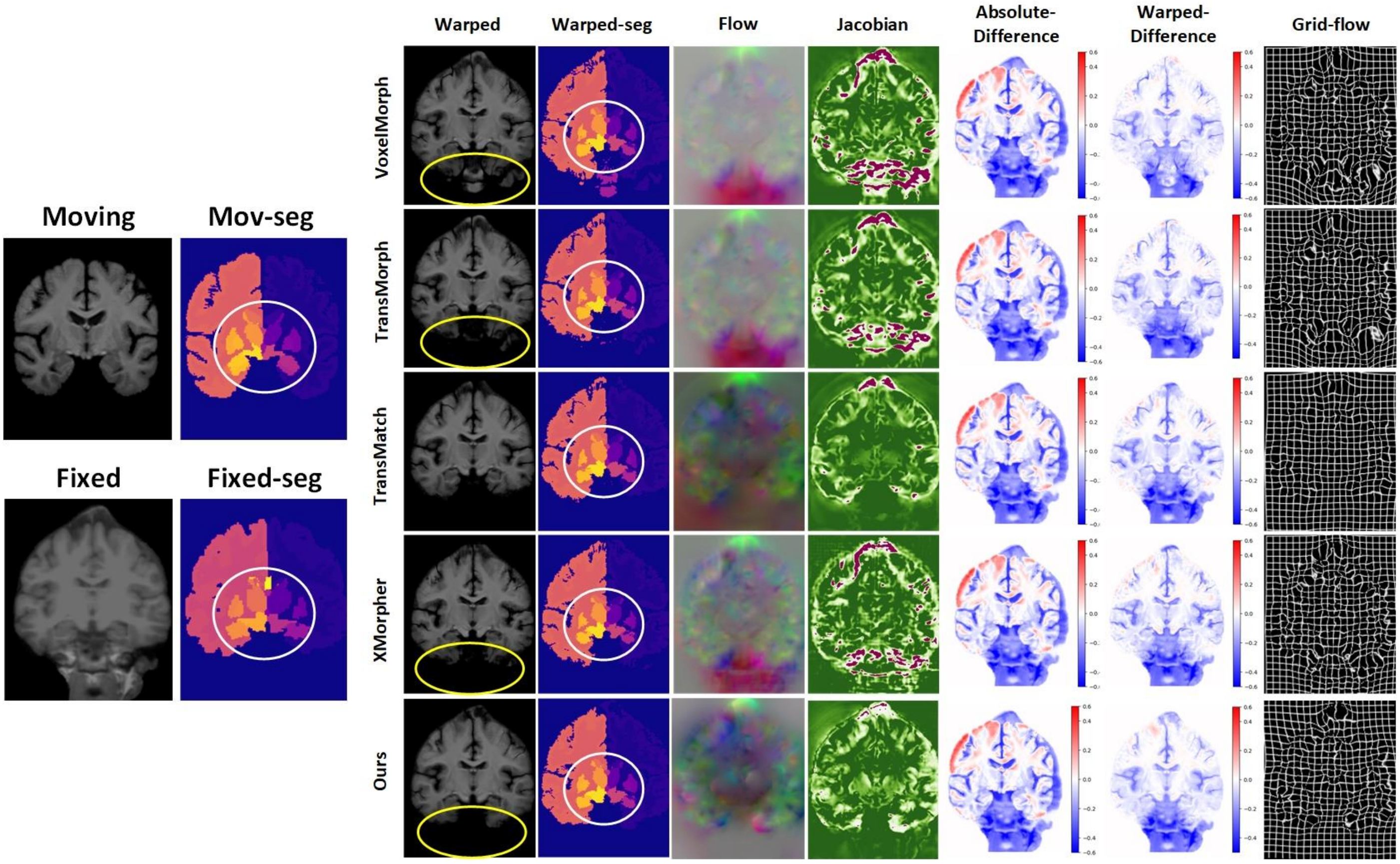


Fig. 7 Visualization of the registration results of the IXI dataset

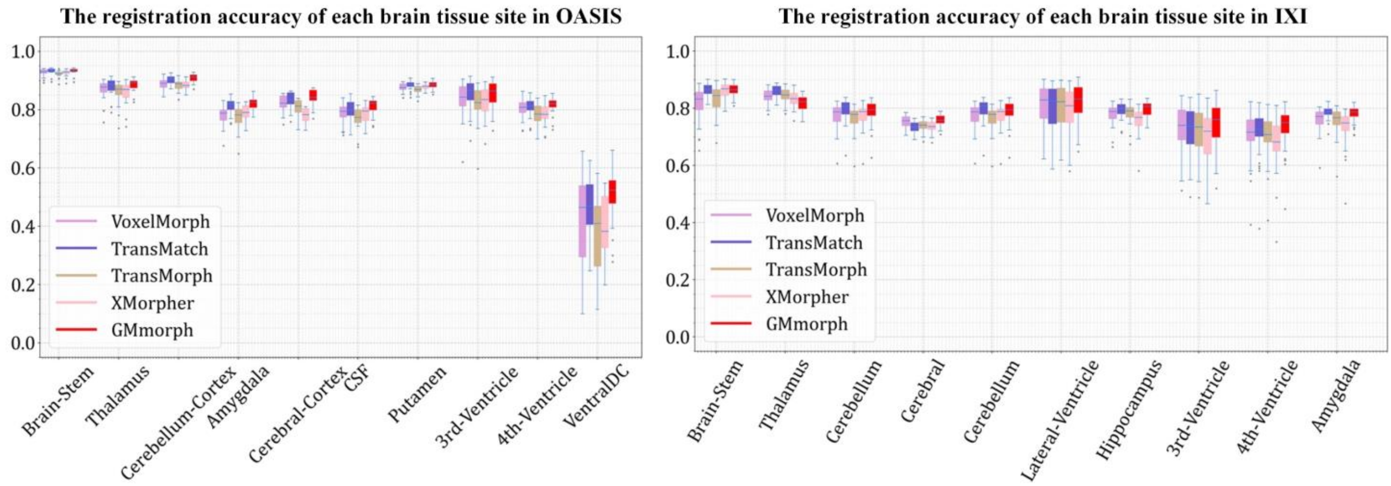
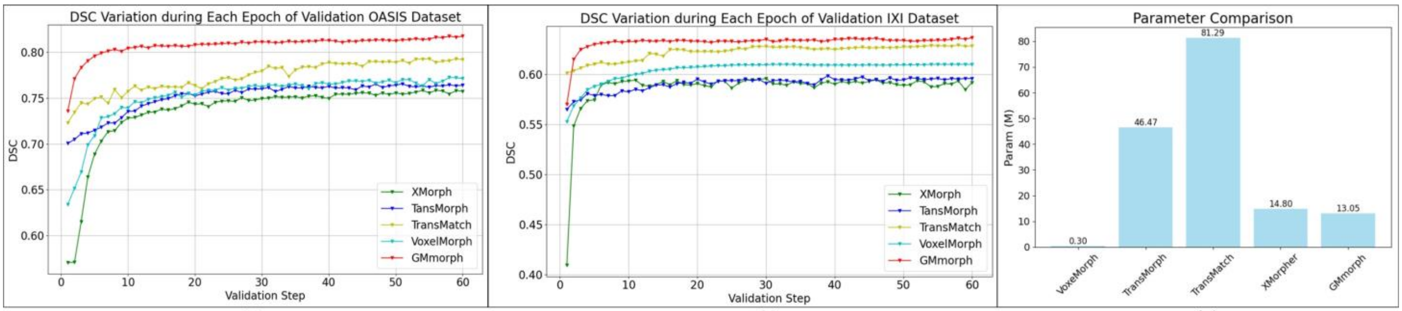


Fig. 8 Box plots of DSC for various comparative models, including10 brain regions such as the brainstem,on the OASIS and IXI dataset.



(a) (b) (c)

Fig. 9 (a) and (b) denote the line plots depicting the change in DSC during the training process for each model. The DSC calculation in this process differs from the testing process in that it selects a larger number of brain tissue regions than the testing process. (c) represents a comparison of the model's parameter count.

3.4 Comparative Methods with Diffeomorphic

We utilized differential isomorphic transformation methods that preserve topology and transform reversibility, as introduced in related works, to design differential registration models in each method. The objective was to observe the reduction in registration accuracy after applying the differential isomorphic operation to the models (Table 3). The settings and implementation details of the relevant hyperparameters were consistent with VoxelMorph-diff, and interested readers can refer to the literature [58] . The DSC difference maps between each model and its corresponding differential isomorphic model are shown in Figure 10. In addition, to visually assess the registration capability of the differential registration models, we present a set of representative registration results in Figure 11.

The differential isomorphic structure ensures zero folding points in the deformation fields of each model, resulting in a decrease in the registration metrics, while there is no significant difference in HD95 values. From Figure 10, it can be observed that Transmatch, Xmorpher, and GMmorph exhibit similar abilities to resist the accuracy decrease caused by differential isomorphic transformations. When observing the visual results in Figure 11, it is evident that the Flow column appears overly smooth and fails to depict contour edge information, resulting in unclear portrayal of tissue details in the registered images. Furthermore, it can be seen from the yellow circle markings that the models do not undergo sufficient deformation, but their Grid-flow performances are excellent.

Table 3: Objective Evaluation Analysis of Differential Isomorphic Structures for Each Model on OASIS

OASIS	**Comprehensive**			**Haus Distance**	**Jacobian<0**
	DSC	**SSIM**	**D-S**		
VoxeMorph_diff	0.751 ±0.025	0.9242±0.0054	0.729±0.019	2.3466±0.4802	0
TransMorph_diff	0.756±0.022	0.9066±0.0057	0.729±0.018	2.3490 ±0.4904	0
TransMatch_diff	0.788±0.025	0.9322+0.006	0.770±0.019	2.0673±0.4646	0
XMorpher_diff	0.765±0.020,	0.8898±0.006	0.734±0.016	2.2679±0.4358	0
GMmorph_diff	0.817±0.024	0.9333±0.0052	0.800±0.019	1.9391±0.4502	0

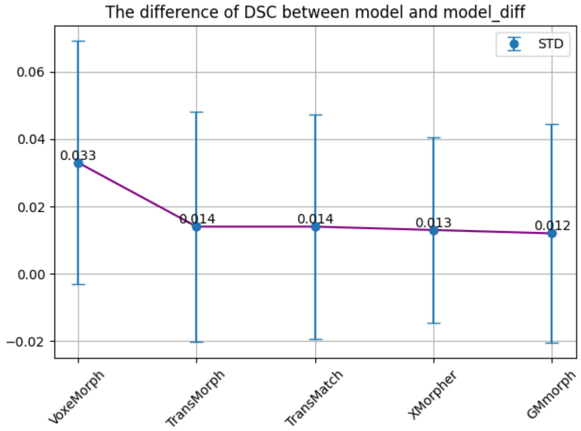


Fig. 10 DSC difference graphs of each model and its corresponding differential homeomorphic model

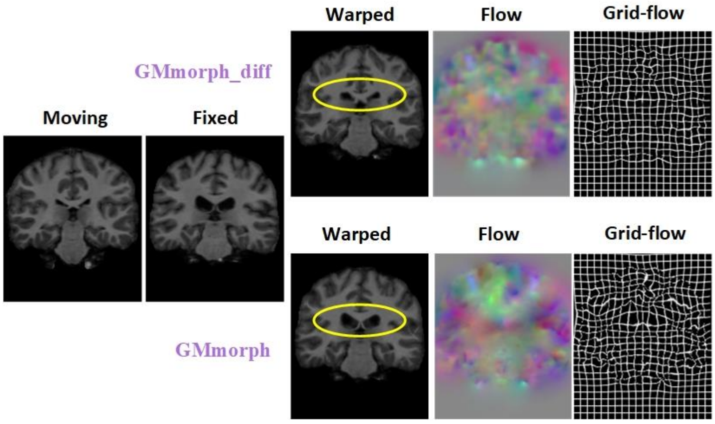


Fig. 11 Visualization of GMmorph_diff and GMmorph registration results

3.5 Ablation results

To validate the effectiveness of each component of the model, we conducted ablation experiments on GMmorph using the IXI dataset. Four ablation schemes were designed: (1) w/o Con-loss: without consistency loss, (2) w/o DMM: without the DMM module, in which the upper and lower branches of the network are connected by a single convolutional layer, (3) w/o GML: without the gated Mamba module, and (4) w/o DEM: without the detail enhancement module. The objective experimental results are shown in Table 4. In order to visually observe the effectiveness of each module in data processing, we present the relevant feature maps in the network structure in Figure 12.

Analyzing Table 4, we can observe that (1) the number of folding points in the model without Con-loss is significantly higher than other schemes, confirming the effectiveness of consistency loss in preserving topological structure and achieving implicit regularization. (2) All ablation schemes show a certain decrease in performance compared to GMmorph, demonstrating the effectiveness of the model's components and losses. (3) The folding phenomenon is least evident when the DEM module is not present, indicating that capturing details and deformations enhances the number of folding points. (4) An increase in accuracy implies an exacerbation of folding phenomena. However, the model is still able to appropriately handle the relationship between the two. Observing the visualized feature maps in Figure 12, the pixel-level features shown in the top-left portion reveal that the presence of the DMM module allows the model to extract more complete pixel-level features. The middle layer features of the network, shown in the top-right portion, clearly demonstrate that the presence of the DMM module improves the activation of region-specific tissue features, with the activated features being more dispersed. The bottom-left portion displays the feature maps near the output layer, where the presence of the DMM module results in smooth and evenly highlighted information. This suggests that the DMM module influences the feature extraction of the upper and lower branches of the model in a spatially matched manner. The learnable offsets enable the model to more accurately match the correct spatial information, exhibiting smoother, more complete, and evenly highlighted information. The bottom-right portion of Figure 12 shows the coarse and fine-grained features learned by the GML and DEM modules. The output of the GML module reveals overall contour information, and upon closer inspection, traces of image segmentation can be observed. The output features of the DEM module are brighter and contain richer information, which confirms that these two modules capture the desired information.

Table 4: Objective evaluation of ablation experiments on IXI

IXI	**Comprehensive**			**Haus Distance**	**Jacobian<0**
	DSC	**SSIM**	**D-S**		
GMmorph	0.764±0.030	0.9323±0.0189	0.745±0.024	5.269±0.8415	26637±8002
w/o Con-loss	0.732±0.032	0.9199±0.0201	0.708±0.025	5.6509±0.7187	81910±15227
w/o DMM	0.744±0.033	0.9058±0.0215	0.716±0.026	5.1725±0.7542	8135±2687
w/o GML	0.746±0.031	0.9207±0.0202	0.723±0.025	5.2690±0.7376	13148±4338
w/o DEM	0.738±0.033	0.8965±0.0219	0.707±0.026	5.23±0.8273	2328±771

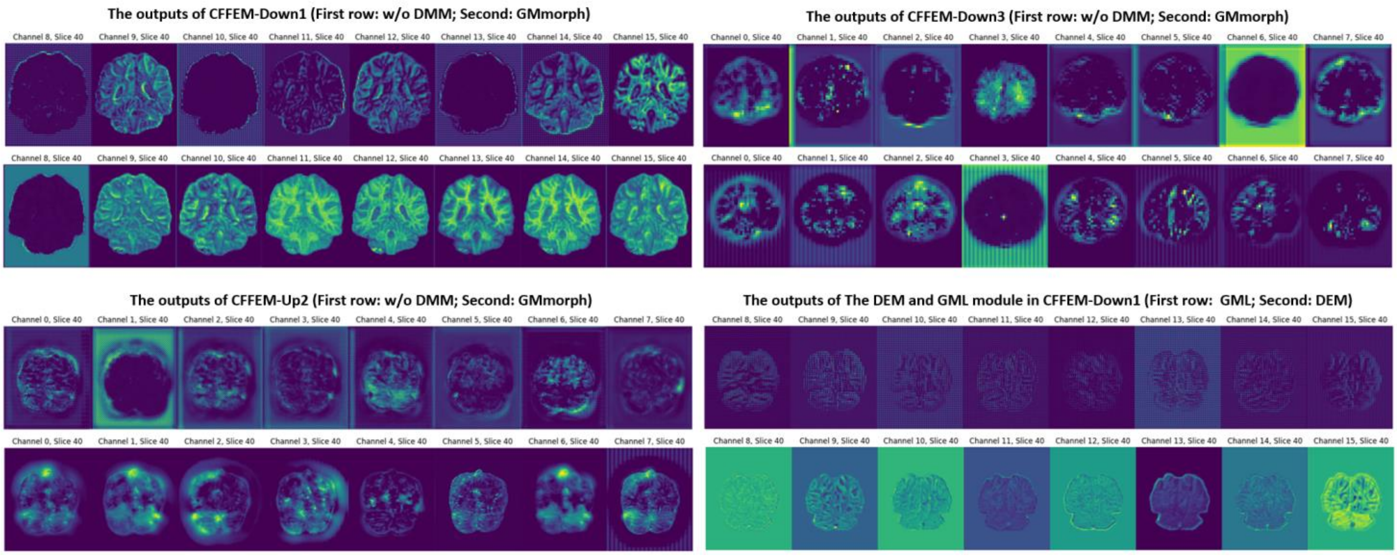


Fig. 12 Visual Representation of Feature Maps for Ablation Experiments on Each Module

3.6 Semi-supervised results

For datasets that contain segmentation labels, these labels are often a better way to facilitate model parameter updates. This can be achieved indirectly by using Dice loss to promote alignment with corresponding anatomical regions or by employing segmentation networks as indirect constraints. In order to further improve the efficiency of semi-supervised registration, we adopted a strategy of cross-inputting labels and data. To verify its effectiveness, we designed a set of comparative experiments in this section, namely GMmorph_S: without cross-inputting, using Dice loss; GMmorph_SC: with cross-inputting, using Dice loss. The experimental results are shown in Table 5, which are also consistent with our expectations. Cross-inputting further enhances the performance of semi-supervised registration, confirming the effectiveness of the aforementioned label-based soft constraints.

Table 5: Evaluation Analysis of Semi-Supervised Registration Schemes on IXI

IXI	Comprehensive			Haus Distance	Jacobian<0
	DSC	SSIM	D-S		
GMmorph	0.764±0.030	0.9323±0.0189	0.745±0.024	5.269±0.8415	26637±8002
GMmorph_S	0.823±0.020	0.9211±0.0197	0.804±0.017	3.7778±0.4859	24355±7456
GMmorph_SC_	0.832±0.016	0.8917±0.0211	0.806±0.014	3.9776±0.6138	17664±6303

3.7 Muti-model registration

In this section, we investigated the effectiveness of GMmorph in multimodal image registration. We selected three modalities of MR images from the Brats2021 dataset: Flair, T1, and T2, and performed registration tasks for T1-T2 and Flair-T1. We did not compare with image transformation-based methods [30, 59] or supervised registration approaches [60, 61] because these methods do not rely on structure or intensity-based metrics to drive the registration process. Additionally, different metrics have variability in registering different modal images. Therefore, we chose NMI and MIND as registration schemes and compared them with VoxelMorph. Objective experimental results are shown in Table 6 and Table 7, and visualizations are presented in Figure 13.

Analysis of the objective results reveals that MIND ensures a lower number of foldings, while NMI encourages the model to produce more diverse displacements. Based on NMI, GMmorph achieves a better balance between accuracy and the number of foldings compared to VoxelMorph, significantly reducing the number of foldings. Overall, GMmorph achieves more optimal and suboptimal results. By observing the difference images in Figure 13, it can be seen that our method provides smoother characterization of edges and clearer overall images. GMmorph exhibits a smoother grid flow. However, in the registration results of Flair-T1 based on NMI and MIND, the selected Flair modality has independent tumor features compared to T1. MIND, which is based on regional self-similarity, can demonstrate more reasonable deformations when dealing with these modality-specific features. It can be observed in the Flair-T1 results under the VM_NMI column that the tumor region almost covers the middle area of the image. In comparison, GMmorph_NMI shows some alleviation capability compared to the former.

Table 6: Multimodal T1-T2 Registration

Brats2021	MI	MAE	Jacobian<0
VoxelMorph_NMI	0.6447±0.0198	0.0847±0.0302	20079±13834
VoxelMorph_MIND	0.6014±0.0415	0.0838±0.0313	2197±5185
GMmorph_NMI	0.6454±0.0194	0.0811±0.0268	3499±5090
GMmorph_MIND	0.6496±0.0175	0.0815±0.0277	166±320

Table 7: Multimodal Flair-T1 Registration

Brats2021	MI	MAE	Jacobian<0
VoxelMorph_NMI	0.6138±0.1044	0.0697±0.0295	19932±11845
VoxelMorph_MIND	0.5658±0.0978	0.0753±0.0291	1260±2861
GMmorph_NMI	0.6097±0.1074	0.0724±0.0302	4935±5050
GMmorph_MIND	0.6010±0.1080	0.0732±0.0303	89±95

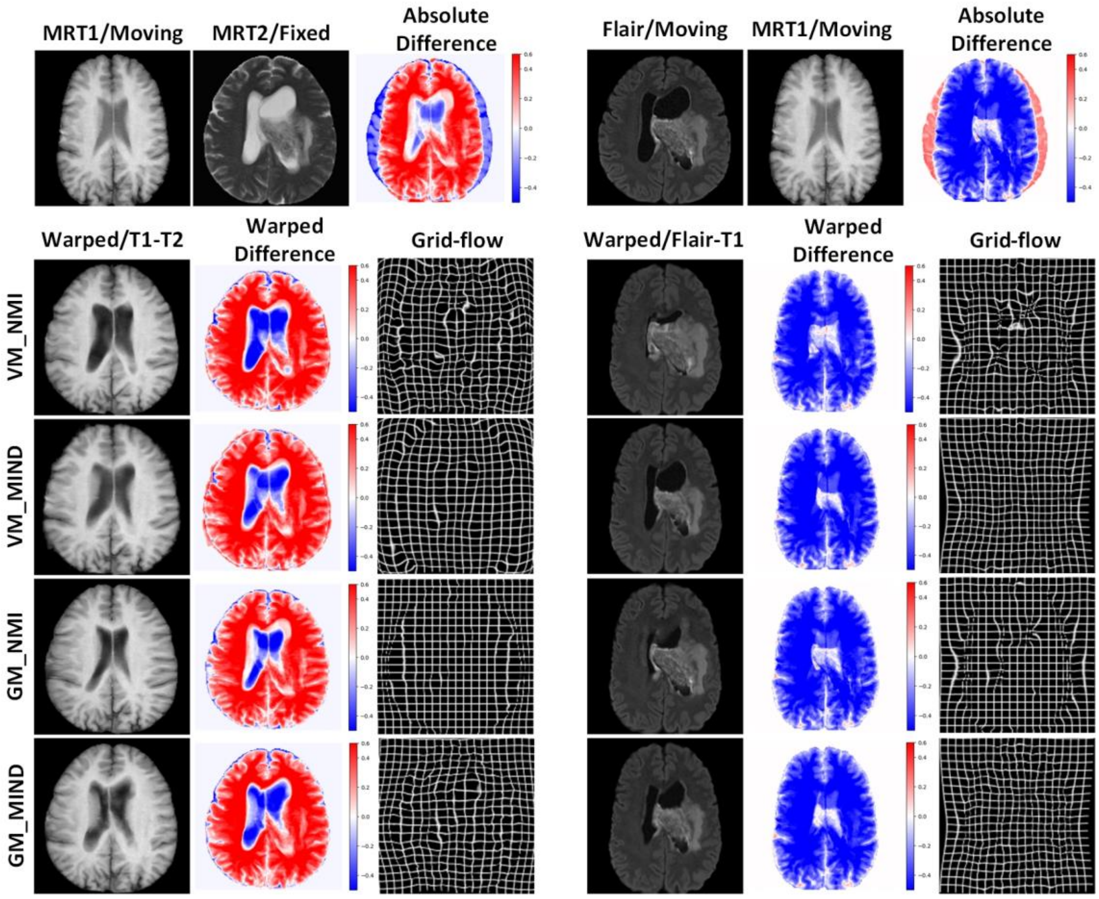


Fig. 13 Visualization of Multimodal Image Registration Results (T1-T2: Registering MRT1 to MRT2; Flair-T1: Registering MR-Flair to MRT1). GM: GMmorph. VM: VoxelMorph.

4. Summary and Discussion

Motivated by the need to enhance the flexibility of spatial matching and the adequacy of feature extraction in the registration process, we propose a dual-branch interactive registration model called GMmorph. GMmorph utilizes the Deformable Matching Mechanism (DMM) module to achieve spatial matching at each network layer using learnable offset-based deformations. This allows the sampling range of the reference image to take any shape based on the offset, enabling the model to find the optimal solution even beyond the sampling range. Additionally, we design a Gate-controlled Mamba Layer (GML) using the Mamba framework to extract global features. Furthermore, we incorporate a Detail Enhancement Module (DEM) based on channel and spatial attention to capture fine details. The model ensures information transfer between the upper and lower branches through the consistency loss, achieving implicit regularization and balancing high registration accuracy with a low number of foldings. We conducted extensive work, including registration tasks on single-modality and multimodal images, evaluation of differential isomorphic registration models, exploration of model robustness and registration performance in the presence of abnormal tissue scenarios, and investigation of the cross-inputting strategy for semi-supervised tasks with image segmentation labels.

GMmorph has achieved outstanding performance in various experiments. However, the model still has some limitations: (1) computational complexity, as the Mamba layer incurs a significant inference cost at the shallow layers of the network, despite achieving linear complexity and a global receptive field. Additionally, the model performs two registration processes in a single inference for the interactive registration structure, providing multiple perspectives to physicians but increasing the inference complexity. (2) The model is not friendly to small datasets, meaning it may not achieve high registration accuracy on such datasets. This is partly related to the modeling mechanism of Mamba, as it requires a certain amount of data to support its global view, similar to VIT.

In future work, we will focus on finding better registration model designs that prioritize practicality and lightweightness. We will explore more effective solutions to address intensity distribution differences and the decoupling of shared and independent features in cross-modal image registration. Additionally, we will strive to design a universal registration model that can generalize across different datasets. We hope that this work will not only contribute to the advancement of the field but also inspire and provide new directions for future registration research.

Reference

[1] G.E. Christensen, R.D. Rabbitt, M.I. Miller, Deformable templates using large deformation kinematics, IEEE transactions on image processing 5(10) (1996) 1435-1447.

[2] M. Cabezas, A. Oliver, X. Lladó, J. Freixenet, M.B. Cuadra, A review of atlas-based segmentation for magnetic resonance brain images, Computer methods and programs in biomedicine 104(3) (2011) e158-e177.

[3] P. Aljabar, R.A. Heckemann, A. Hammers, J.V. Hajnal, D. Rueckert, Multi-atlas based segmentation of brain images: atlas selection and its effect on accuracy, Neuroimage 46(3) (2009) 726-738.

[4] R. Shams, P. Sadeghi, R.A. Kennedy, R.I. Hartley, A survey of medical image registration on multicore and the GPU, IEEE signal processing magazine 27(2) (2010) 50-60.

[5] B.B. Avants, C.L. Epstein, M. Grossman, J.C. Gee, Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain, Medical image analysis 12(1) (2008) 26-41.

[6] S. Klein, M. Staring, K. Murphy, M.A. Viergever, J.P. Pluim, Elastix: a toolbox for intensity-based medical image registration, IEEE transactions on medical imaging 29(1) (2009) 196-205.

[7] S. Gold, C.-P. Lu, A. Rangarajan, S. Pappu, E. Mjolsness, New algorithms for 2D and 3D point matching: Pose estimation and correspondence, Advances in neural information processing systems 7 (1994).

[8] D. Shen, C. Davatzikos, HAMMER: hierarchical attribute matching mechanism for elastic registration, IEEE transactions on medical imaging 21(11) (2002) 1421-1439.

[9] B.S. Reddy, B.N. Chatterji, An FFT-based technique for translation, rotation, and scale-invariant image registration, IEEE transactions on image processing 5(8) (1996) 1266-1271.

[10] F.P. Oliveira, J.M.R. Tavares, Medical image registration: a review, Computer methods in biomechanics and biomedical engineering 17(2) (2014) 73-93.

[11] M.A. Viergever, J.A. Maintz, S. Klein, K. Murphy, M. Staring, J.P. Pluim, A survey of medical image registration - under review, Elsevier, 2016, pp. 140-144.

[12] C. Davatzikos, Spatial transformation and registration of brain images using elastically deformable models, Computer Vision and Image Understanding 66(2) (1997) 207-222.

[13] R. Bajcsy, S. Kovačič, Multiresolution elastic matching, Computer vision, graphics, and image processing 46(1) (1989) 1-21.

[14] Y. Gal, Z. Ghahramani, Dropout as a bayesian approximation: Representing model uncertainty in deep learning, international conference on machine learning, PMLR, 2016, pp. 1050-1059.

[15] M.-H. Laves, M. Tölle, T. Ortmaier, Uncertainty estimation in medical image denoising with bayesian deep image prior, Uncertainty for Safe Utilization of Machine Learning in Medical Imaging, and Graphs in Biomedical Image Analysis: Second International Workshop, UNSURE 2020, and Third International Workshop, GRAIL 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 8, 2020, Proceedings 2, Springer,

2020, pp. 81-96.

[16] P. Risholm, F. Janoos, I. Norton, A.J. Golby, W.M. Wells III, Bayesian characterization of uncertainty in intra-subject non-rigid registration, Medical image analysis 17(5) (2013) 538-555.

[17] M.F. Beg, M.I. Miller, A. Trouvé, L. Younes, Computing large deformation metric mappings via geodesic flows of diffeomorphisms, International journal of computer vision 61 (2005) 139-157.

[18] T. Vercauteren, X. Pennec, A. Perchant, N. Ayache, Diffeomorphic demons: Efficient non-parametric image registration, NeuroImage 45(1) (2009) S61-S72.

[19] A. Sotiras, C. Davatzikos, N. Paragios, Deformable medical image registration: A survey, IEEE transactions on medical imaging 32(7) (2013) 1153-1190.

[20] M. Jaderberg, K. Simonyan, A. Zisserman, K. Kavukcuoglu, Spatial transformer networks. Advances in Neural Information Processing Systems 28, Annual Conference on Neural Information Processing Systems, 2015, pp. 7-12.

[21] M.-M. Rohé, M. Datar, T. Heimann, M. Sermesant, X. Pennec, SVF-Net: learning deformable image registration using shape matching, Medical Image Computing and Computer Assisted Intervention− MICCAI 2017: 20th International Conference, Quebec City, QC, Canada, September 11-13, 2017, Proceedings, Part I 20, Springer, 2017, pp. 266-274.

[22] S.S.M. Salehi, S. Khan, D. Erdogmus, A. Gholipour, Real-time deep pose estimation with geodesic loss for image-to-template rigid registration, IEEE transactions on medical imaging 38(2) (2018) 470-481.

[23] J. Fan, X. Cao, P.-T. Yap, D. Shen, BIRNet: Brain image registration using dual-supervised fully convolutional networks, Medical image analysis 54 (2019) 193-206.

[24] S. Miao, Z.J. Wang, R. Liao, A CNN regression approach for real-time 2D/3D registration, IEEE transactions on medical imaging 35(5) (2016) 1352-1363.

[25] Y. Hu, M. Modat, E. Gibson, W. Li, N. Ghavami, E. Bonmati, G. Wang, S. Bandula, C.M. Moore, M. Emberton, Weakly-supervised convolutional neural networks for multimodal image registration, Medical image analysis 49 (2018) 1-13.

[26] X. Cao, J. Yang, J. Zhang, D. Nie, M. Kim, Q. Wang, D. Shen, Deformable image registration based on similarity-steered CNN regression, Medical Image Computing and Computer Assisted Intervention− MICCAI 2017: 20th International Conference, Quebec City, QC, Canada, September 11-13, 2017, Proceedings, Part I 20, Springer, 2017, pp. 300-308.

[27] X. Yang, R. Kwitt, M. Styner, M. Niethammer, Quicksilver: Fast predictive image registration–a deep learning approach, NeuroImage 158 (2017) 378-396.

[28] J. Chen, Y. Liu, S. Wei, Z. Bian, S. Subramanian, A. Carass, J.L. Prince, Y. Du, A Survey on Deep Learning in Medical Image Registration: New Technologies, Uncertainty, Evaluation Metrics, and Beyond, arXiv preprint arXiv:2307.15615 (2023).

[29] J. Chen, E.C. Frey, Y. He, W.P. Segars, Y. Li, Y. Du, Transmorph: Transformer for unsupervised medical image registration, Medical image analysis 82 (2022) 102615.

[30] A. Casamitjana, M. Mancini, J.E. Iglesias, Synth-by-reg (sbr): Contrastive learning for synthesis-based registration of paired images, Simulation and Synthesis in Medical Imaging: 6th International Workshop, SASHIMI 2021, Held in Conjunction with MICCAI 2021, Strasbourg, France, September 27, 2021, Proceedings 6, Springer, 2021, pp. 44-54.

[31] B. Kim, D.H. Kim, S.H. Park, J. Kim, J.-G. Lee, J.C. Ye, CycleMorph: cycle consistent unsupervised deformable image registration, Medical image analysis 71 (2021) 102036.

[32] S. Zhao, Y. Dong, E.I. Chang, Y. Xu, Recursive cascaded networks for unsupervised medical image registration, Proceedings of the IEEE/CVF international conference on computer vision, 2019, pp. 10600-10610.

[33] M. Meng, L. Bi, M. Fulham, D. Feng, J. Kim, Non-iterative Coarse-to-Fine Transformer Networks for Joint Affine and Deformable Image Registration, International Conference on Medical Image Computing and

Computer-Assisted Intervention, Springer, 2023, pp. 750-760.

[34] Y. Wang, W. Qian, M. Li, X. Zhang, A transformer-based network for deformable medical image registration, CAAI International Conference on Artificial Intelligence, Springer, 2022, pp. 502-513.

[35] X. Song, H. Chao, X. Xu, H. Guo, S. Xu, B. Turkbey, B.J. Wood, T. Sanford, G. Wang, P. Yan, Cross-modal attention for multi-modal image registration, Medical Image Analysis 82 (2022) 102612.

[36] J. Shi, Y. He, Y. Kong, J.-L. Coatrieux, H. Shu, G. Yang, S. Li, Xmorpher: Full transformer for deformable medical image registration via cross attention, International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2022, pp. 217-226.

[37] Z. Chen, Y. Zheng, J.C. Gee, TransMatch: A Transformer-based Multilevel Dual-Stream Feature Matching Network for Unsupervised Deformable Image Registration, IEEE Transactions on Medical Imaging (2023).

[38] J. Zhang, T. Fu, J. Li, D. Xiao, J. Fan, Y. Lin, H. Song, F. Ji, M. Yang, J. Yang, An alternately optimized generative adversarial network with texture and content constraints for deformable registration of 3D ultrasound images, Physics in Medicine and Biology (2023).

[39] X. Zhu, M. Ding, X. Zhang, Free form deformation and symmetry constraint‐based multi‐modal brain image registration using generative adversarial nets, CAAI Transactions on Intelligence Technology (2023).

[40] X. Fan, Z. Li, Z. Li, X. Wang, R. Liu, Z. Luo, H. Huang, Automated learning for deformable medical image registration by jointly optimizing network architectures and objective functions, IEEE Transactions on Image Processing (2023).

[41] E. Pachetti, S. Colantonio, A Systematic Review of Few-Shot Learning in Medical Imaging, arXiv preprint arXiv:2309.11433 (2023).

[42] G. Balakrishnan, A. Zhao, M.R. Sabuncu, J. Guttag, A.V. Dalca, VoxelMorph: a learning framework for deformable medical image registration, IEEE transactions on medical imaging 38(8) (2019) 1788-1800.

[43] Z. Chen, E.N. Brown, State space model, Scholarpedia 8(3) (2013) 30868.

[44] A. Gu, T. Dao, Mamba: Linear-time sequence modeling with selective state spaces, arXiv preprint arXiv:2312.00752 (2023).

[45] L. Zhu, B. Liao, Q. Zhang, X. Wang, W. Liu, X. Wang, Vision mamba: Efficient visual representation learning with bidirectional state space model, arXiv preprint arXiv:2401.09417 (2024).

[46] Y. Liu, Y. Tian, Y. Zhao, H. Yu, L. Xie, Y. Wang, Q. Ye, Y. Liu, Vmamba: Visual state space model, arXiv preprint arXiv:2401.10166 (2024).

[47] A. Gu, Modeling Sequences with Structured State Spaces, Stanford University2023.

[48] A. Gu, K. Goel, C. Ré, Efficiently modeling long sequences with structured state spaces, arXiv preprint arXiv:2111.00396 (2021).

[49] A. Gu, T. Dao, S. Ermon, A. Rudra, C. Ré, Hippo: Recurrent memory with optimal polynomial projections, Advances in neural information processing systems 33 (2020) 1474-1487.

[50] J. Ma, F. Li, B. Wang, U-mamba: Enhancing long-range dependency for biomedical image segmentation, arXiv preprint arXiv:2401.04722 (2024).

[51] M.P. Heinrich, M. Jenkinson, M. Bhushan, T. Matin, F.V. Gleeson, M. Brady, J.A. Schnabel, MIND: Modality independent neighbourhood descriptor for multi-modal deformable registration, Medical image analysis 16(7) (2012) 1423-1435.

[52] P. Viola, W.M. Wells III, Alignment by maximization of mutual information, International journal of computer vision 24(2) (1997) 137-154.

[53] A. Hoopes, M. Hoffmann, B. Fischl, J. Guttag, A.V. Dalca, Hypermorph: Amortized hyperparameter learning for image registration, Information Processing in Medical Imaging: 27th International Conference, IPMI 2021, Virtual Event, June 28–June 30, 2021, Proceedings 27, Springer, 2021, pp. 3-17.

[54] B. Fischl, FreeSurfer, Neuroimage 62(2) (2012) 774-781.

[55] A. Hore, D. Ziou, Image quality metrics: PSNR vs. SSIM, 2010 20th international conference on pattern recognition, IEEE, 2010, pp. 2366-2369.

[56] D.P. Huttenlocher, G.A. Klanderman, W.J. Rucklidge, Comparing images using the Hausdorff distance, IEEE Transactions on pattern analysis and machine intelligence 15(9) (1993) 850-863.

[57] Z. Yu, L. Chen, Z. Cheng, J. Luo, Transmatch: A transfer-learning scheme for semi-supervised few-shot learning, Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 12856-12864.

[58] A.V. Dalca, G. Balakrishnan, J. Guttag, M.R. Sabuncu, Unsupervised learning of probabilistic diffeomorphic registration for images and surfaces, Medical image analysis 57 (2019) 226-236.

[59] C. Lian, X. Li, L. Kong, J. Wang, W. Zhang, X. Huang, L. Wang, CoCycleReg: Collaborative cycle-consistency method for multi-modal medical image registration, Neurocomputing 500 (2022) 799-808.

[60] T. Guo, Y. Wang, C. Meng, Mambamorph: a mamba-based backbone with contrastive feature learning for deformable mr-ct registration, arXiv preprint arXiv:2401.13934 (2024).

[61] X. Deng, E. Liu, S. Li, Y. Duan, M. Xu, Interpretable multi-modal image registration network based on disentangled convolutional sparse coding, IEEE Transactions on Image Processing 32 (2023) 1078-1091.

Cover letter

Dear Editor:

It is with great honor and joy that we submit our work to your esteemed journal. After a long period of effort, we have completed this registration work related to Mamba.

This work concerns 3D deformable medical image registration, addressing the problem of low registration accuracy caused by the lack of flexibility in previous methods and insufficient feature extraction in the spatial matching process. **(1)** Based on Mamba, we designed the Gated Mamba layer to achieve global feature extraction, combined with attention mechanism to design detail enhancement modules. **(2)** The interactive registration and deformable dynamic matching mechanism (DMM) promote the model to complete a more accurate spatial matching process. **(3)** We evaluated the model on tasks of multi-modal and single-modal image registration, including semi-supervised and unsupervised modes, and in complex registration scenarios. **(4)** This article has achieved very good accuracy. The pattern of interactive registration combined with the constraint of consistency loss can effectively enhance the robustness of the model.

The linear complexity and global receptive field of Mamba enable it to operate in the shallow layers of the network, which is highly suitable for registration. This article has been tested on various datasets, and we believe that this work can make some contributions to this field.

If you have any questions, please feel free to contact us. Wishing you a pleasant life and smooth work!

Yours,
Hao Lin
2024/4/28

Title: GMmorph: Dynamic Spatial Matching Registration Model for 3D Medical Image based on Gated Mamba

Authors:

First author: Hao Lin (PhD candidate, School of Software, Xi 'an Jiaotong University, Xi'an City, Shanxi Province, China)   Mail: 4123158007@stu.xjtu.edu.cn

Second author and Corresponding author: Yonghong Song (Professor, School of Software, Xi 'an Jiaotong University, Xi'an City, Shanxi Province, China) Mail: songyh@mail.xjtu.edu.cn

Other authors: Fei Li; Qi Zhang; Bincheng Peng.

Post Code: 710049

Author statement:

Hao lin:

Conceptualization; Data curation; Formal analysis; Investigation; Methodology; Software; Validation; Visualization; Roles/Writing - original draft; Writing - review & editing;

Yonghong Song:

Funding acquisition; Writing - review & editing; Supervision; Resources; Project administration;

Others:

Data curation; Formal analysis; Investigation; Validation;

Funding:

This work is supported by the National Natural Science Foundation of China, Grant Number : 62350083 and the Key R&D Plan of Shaanxi Province,Program No. 2023-YBGY-029

Competing Interests:

The authors have no competing interests to declare that are relevant to the content of this article.

Data Availability Statement：

The datasets generated during and/or analysed during the current study are available from the corresponding author on reasonable request.

Abstract:

Medical image registration plays a crucial role in precise planning for radiotherapy, as well as in tasks such as image fusion and medical image analysis. Deep learning registration methods have been widely employed due to their excellent registration efficiency. However, most existing models exhibit limitations in terms of spatial matching and anatomical feature extraction between image pairs, resulting in insufficient registration accuracy and robustness towards abnormal tissues. To address this issue, we propose a dual-branch registration model architecture from the perspectives of spatial matching and feature extraction. Specifically, our approach includes: (1) Designing the Dynamic Matching module (DMM) to perform deformable and dynamic spatial matching, which generates learnable offsets based on real-time information flow in the upper and lower branches, thus promoting flexibility in searching for optimal positions; (2) Introducing the Gated Mamba Layer (GML) to achieve global feature extraction, and using channel and spatial attention to design the Detailed Enhancement Module (DEM) for enhancing local detailed features. The combination of these two modules forms the Coarse-to-Fine Feature Extraction Module (CFFEM), enabling the retrieval of coarse and fine features. Furthermore, an implicit regularization is achieved through a consistency loss, which encourages the network to balance high accuracy, low folding rate, and robustness. Our model demonstrates excellent performance in both single-modal and multi-modal image registration in semi-supervised and unsupervised modes, as validated by comparisons with various state-of-the-art methods on the IXI, OASIS, and Brats2021 datasets.

Author Agreement

Dear Editor:

We the undersigned declare that this manuscript entitled “GMmorph: Dynamic Spatial Matching Registration Model for 3D Medical Image based on Gated Mamba” is original, has not been published before and is not currently being considered for publication elsewhere.We confirm that the manuscript has been read and approved by all named authors and that there are no other persons who satisfied the criteria for authorship but are not listed. We further confirm that the order of authors listed in the manuscript has been approved by all of us.We understand that the Corresponding Author is the sole contact for the Editorial process. She is responsible for communicating with the other authors about progress, submissions of revisions and final approval of proofs.

Signed by all authors as follows:

Hao Lin; Yonghong Song; Fei Li; Qi Zhang; Bincheng Peng.

The authors have no competing interests to declare that are relevant to the content of this article.

Yours,
Hao Lin
2024/4/28