

LLaMA-Reg: Using LLaMA 2 for Unsupervised Medical Image Registration

Mingrui Ma^{a,1}, Yu Yang^{b,c,d,1,*}

^a*Faculty of Information Science Engineering and Automation, Kunming University of Science and Technology, Kunming, China*

^b*Department of Orthopaedics, Taizhou Hospital of Zhejiang Province Affiliated with Wenzhou Medical University, Wenzhou, China*

^c*Department of Orthopaedics, Enze Hospital, Taizhou Enze Medical Centre (Group), Taizhou, China*

^d*Department of Orthopaedics, The Second Affiliated Hospital of Dalian Medical University, Dalian, China*

Abstract

Medical image registration is an essential topic in medical image analysis. In this paper, we propose a method for medical image registration using a pretrained large language model. We find that using the pretrained large language model to encode deep features of the medical images in the registration model can effectively improve image registration accuracy, indicating the great potential of the large language model in medical image registration tasks. We use dual encoders to perform deep feature extraction on image pairs and then input the features into the pretrained large language model. To adapt the large language model to our registration task, the weights of the large language model are frozen in the registration model, and an adapter is utilized to fine-tune the large language model, which aims at (a) mapping the

*Corresponding author.

Email addresses: mamr@kust.edu.cn (Mingrui Ma), yangy6874@enzemed.com (Yu Yang)

¹These authors contributed equally to this work.

visual tokens to the language space before the large language model computing, (b) project the modeled language tokens output from the large language model to the visual space. Our method combines output features from the fine-tuned large language model with the features output from each encoder layer to gradually generate the deformation fields required for registration in the decoder. To demonstrate the effectiveness of the large prediction model in registration tasks, we conducted experiments on knee and brain MRI and achieved state-of-the-art results.

Keywords: Medical image registration, Large language model, LLaMA, Adapter, Deep learning

1. Introduction

Medical image registration plays a crucial role in the field of medical image analysis by seeking to establish a meaningful connection between the voxels within two images. This process has found extensive applications in various domains, including but not limited to muscle segmentation [1], intraoperative localization [2], and quantified ablation margins and local disease progression after thermal ablation of colorectal liver metastases [3]. By focusing on the alignment of images, medical image registration enables the exploration of changes in anatomical structures over time, providing valuable insights into patterns of variation and development.

Over the past decade, Convolutional Neural Networks (CNNs) have made significant strides in computer vision (CV), marked by their notable successes in diverse tasks [4, 5]. Building on this progress and the rapid evolution of CNNs, there has been a distinct shift towards CNN-based approaches

in medical image analysis [6, 7]. The emergence of U-Net and its variants in 2015 underscored its ability to effectively integrate both low-level and high-level semantic information while maintaining a constrained parameter count, leading to widespread adoption in various medical image analysis tasks [8, 9]. It is also employed as the backbone of the registration models, which predict the deformation fields. Driven by the long-range modeling capabilities of Vision Transformers (ViTs), recent registration researches [10, 11, 12] have employed ViT blocks to leverage the modeling power. Besides using powerful modeling blocks, some methods, such as [13, 14], have applied the multi-scale architecture and predicted the deformation fields at different scale stages, where the previously predicted deformation fields control the subsequent generation of the deformation fields. Unlike the multi-scale framework, the cascaded registration approaches [15, 16] have utilized a progressive manner in which a registration sub-model predicts the deformation fields based on the previous sub-model at the same resolution stage.

With the recent emergence of ChatGPT, large models have attracted widespread attention. Large Language Models (LLMs), such as LLaMA 2 [17], Pythia [18], and GPT-3 [19] etc., which had the vast number of parameters, were trained on large-scale corpus data, aiming to understand complex linguistic content to enhance the capabilities of generating appropriate text and chatting to assist people. Recent studies [20, 21, 22] have demonstrated that pretrained LLMs are important for the cross-model (i.e., vision-language) task. The idea of the adapter appeared in the field of domain adaptation [23]. In order to use the powerful modeling capabilities of the pretrained model in downstream tasks, [24] first proposed using adapters

to fine-tune and apply BERT [25] in various text classification tasks. The current LLMs are employed as sub-models in visual-language tasks to generate specific text content. Generally, to align the features from the image and language domains for downstream tasks, some methods [26, 27] utilized linear projections, i.e., adapters, to achieve the alignment of features in these two domains.

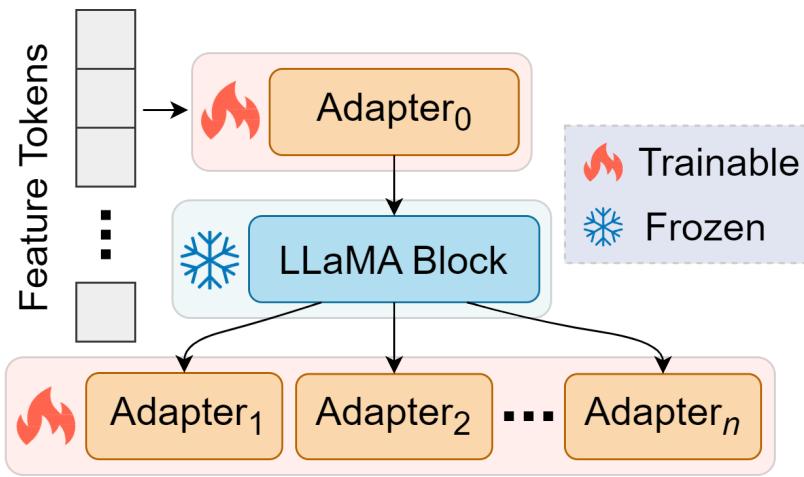


Figure 1: Overview of the utilization of LLaMA in our work. The fire icon represents the trainable block, and the snow icon indicates the frozen LLaMA block with the pretrained weights.

In this paper, we propose a non-U-shaped image registration method, dubbed LLaMA-Reg, which integrates LLaMA 2 as the deep encoder for feature extraction. Our method extracts the deep features of an image pair using a two-stream CNN encoder. These features are split into feature tokens and then sent to the LLaMA encoder. The pretrained LLaMA model and its adapters are shown in Fig. 1. The LLaMA encoder of our method is frozen, which aims to boost the registration performance by using the

projected features in the language domain. Before the LLaMA encoder, an adapter is utilized to project the deep features of two images to the language domain for the pretrained LLaMA encoding, as some other adapters are utilized to project the features in the language domain to the visual domain. The projected features in the visual domain are split into several branches representing different registration stages. We apply the multi-scale framework in each registration stage to achieve coarse-to-fine registration. It is worth noting that the LLaMA block in this figure is an overview. The detail of the designed LLaMA module can be seen in Sec. 3.3.

To evaluate the performance of LLaMA-Reg, we conducted registration tasks on two anatomical image datasets: knee MR image and brain MR image. The quantitative and qualitative comparison results demonstrate the superior performance of LLaMA-Reg over state-of-the-art methods.

The summary contributions of our LLaMA-Reg are as follows:

- We propose an architecture for unsupervised image registration, LLaMA-Reg, which introduces pretrained LLaMA to the registration task and employs a proposed LLaMA block to boost image registration performance.
- We propose a scheme for adapters in the registration task to align the language and visual domain features.
- We propose a non-U-shaped multi-scale and cascaded registration model to utilize LLaMA 2 for registration.
- The experimental results of our LLaMA-Reg on the 3D knee and brain MRI datasets demonstrate state-of-the-art performance. Ablation stud-

ies demonstrate our effectiveness in this work.

2. Related Work

Traditional image registration methods, such as [28], [29], and [30], used iterative optimization to find the best deformation between a pair of images. However, the problem with traditional methods is that they occupy many computing resources and are time-consuming for inference. Methods based on deep learning utilize similarity loss functions to train their weights, which can calculate the deformation field between a pair of images in a very short time after training. Unsupervised deep learning-based registration approaches have been brought to the fore since they do not require ground-truth deformation fields to train. Currently, registration methods based on deep learning that have been widely studied can be divided into two major categories: CNN-based and ViT-based.

2.1. Convolutional Neural Networks-based Approaches

Since the development of CNN-based registration methods, many approaches have emerged. Balakrishnan et al. first introduced VoxelMorph [31], a U-shaped structure registration model for an input pair of images that predicted full-scale displacement fields. In order to ensure the diffeomorphism characteristics of the deformation field between a pair of images and make the deformation smooth, Dalca et al. developed a diffeomorphic registration model [32]. To guarantee some other properties of deformation in registration, Mok et al. proposed SYMNet [33] to predict the bidirectional diffeomorphic deformation field. Kim et al. [34] introduced the cycle consistency registration model to enhance performance and preserve topology.

To further improve the performance of the registration model, Mok et al. proposed a multi-scale registration model [14] based on the Laplacian pyramid network. The multi-scale model performed a coarse-to-fine registration based on the deformation field predicted by the previous scale model. [35] presented a single-pass model that integrated the multi-scale scheme in the decoder to perform coarse-to-fine registration. Cascaded models, such as [15], employed several models to warp moving images gradually.

2.2. *Vision Transformers-based Approaches*

Since the advent of ViT, where Transformer was applied in computer vision, it has attracted the attention of many scholars in the field of medical image analysis. Chen et al. [36] integrated ViT into V-Net at the bottom of their model to perform image registration. Zhang et al. [37] introduced a dual ViT-based network to enhance the feature modeling capability. Ma et al. [11] presented a symmetric variant ViT-based U-Net to improve the registration performance. Chen et al. [10] developed TransMorph, consisting of a Swin transformer-based encoder and a CNN-based decoder. Zhu et al. [38] designed a symmetric Swin transformer-based architecture that maintains invertibility and topology preservation. In order to solve the problem that the transformation of image features to image matching relationships is implicit, TranMatch [39], a dual stream feature matching registration model, was proposed based on the Swin Transformer. All these ViT-based approaches mentioned above indicated performance improvement benefiting from the strong modeling power of ViTs.

3. Methods

3.1. Image Registration

Learning-based deformable image registration minimizes a similarity energy function to establish a dense spatial correspondence between an image pair. Given an image pair $\{I_m, I_f\}$ defined on a 3D domain $\Omega \subset \mathbb{R}^3$ denoting moving and fixed images. Optimization aims to find an optimal deformation field that can be formulated as

$$\hat{\phi} = \arg \min_{\phi} (\mathcal{L}_{sim}(I_m \circ \phi, I_f) + \lambda \mathcal{L}_{reg}(\phi)), \quad (1)$$

where the I_m and I_f denote the moving and fixed image, $I_m \circ \phi$ is the warped moving image transformed via a deformation field ϕ . \mathcal{L}_{sim} is the similarity matrix to estimate the similarity between $I_m \circ \phi$ and I_f . $\mathcal{L}_{reg}(\phi)$ is the regularization, which enforces the smoothness of the deformation field by the spatial gradient, and λ is a hyperparameter used to balance contribution in the learning of similarity and smoothness. Hence, the optimal deformation field $\hat{\phi}$ is obtained.

In this work, we follow Eq. 1 to train our deformable image registration model unsupervised. Mean squared error (MSE) is utilized as the similarity metric to evaluate the similarity between an image pair, i.e., $\mathcal{L}_{sim} = \text{MSE}(I_m \circ \phi, I_f)$. \circ is the spatial transform network (STN) [40], and $I_m \circ \phi$ represents I_m warped via a deformation field ϕ . STN can warp an image with a deformation field in an interpolation manner. We utilize the diffusion regularizer [41] on the spatial gradients of a deformation field ϕ . The regularizer is denoted as $\mathcal{L}_{reg} = \text{Diff}(\phi)$. Hence, the loss function in this work is $\mathcal{L}(I_m, I_f, \phi) = \text{MSE}(I_m \circ \phi, I_f) + \lambda \text{Diff}(\phi)$, where λ is the hyperparameter

that determines the trade-off between similarity and regularity. We optimize the parameters of LLaMA-Reg by minimizing this loss function.

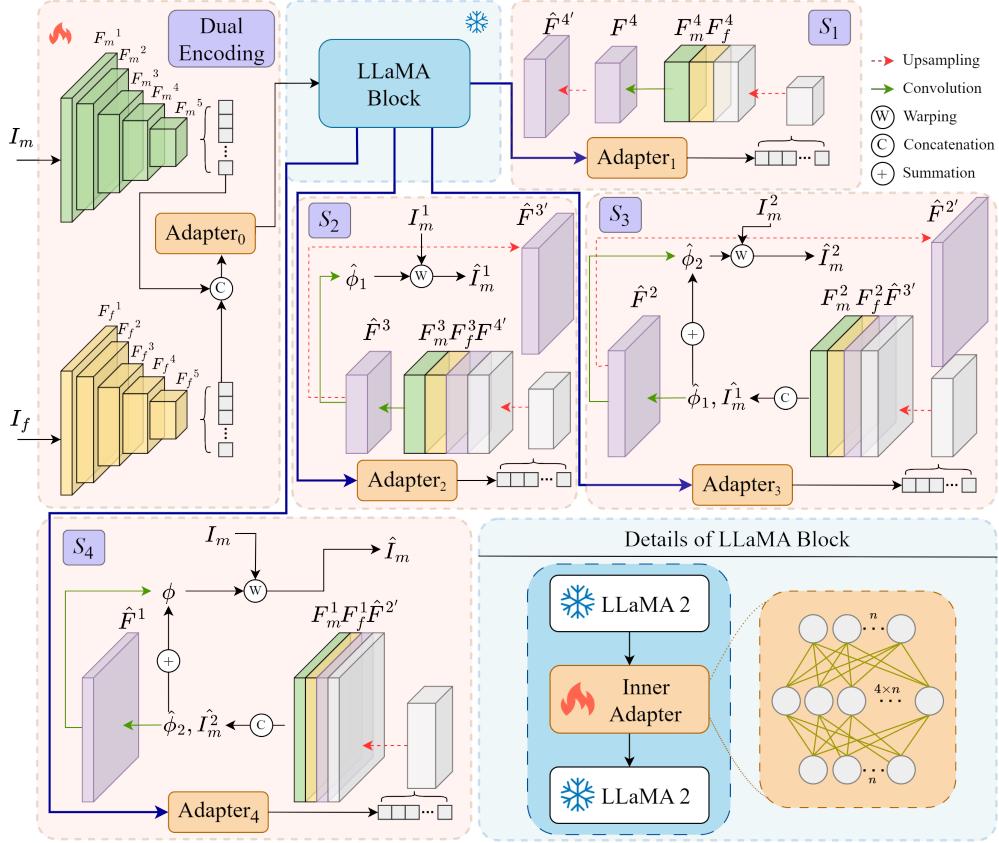


Figure 2: The overview of our proposed LLaMA-Reg. LLaMA-Reg employs a dual encoding block to extract the deep features of I_m and I_f . Adapters are utilized to project visual features to the language domain or language features to the visual domain. S_1 , S_2 , S_3 , and S_4 represent layers at different resolution stages from 1/8 to full resolution scale. Our model performs image registration at these stages to achieve a multi-scale registration manner. The LLaMA block is shown in the light blue box, and we offer its details in the lower right.

3.2. Adapters for Cross-Domain Projection

For the fine-tuning methods, adapter tuning is an efficient manner to fine-tune the pretrained model with few parameters gained. In our work, we borrow the idea of adapter-tuning and use adapters composed of linear projections outside LLaMA so that the visual features obtained before LLaMA can be fitted to the parameter space (i.e., language domain) of LLaMA. More specifically, the extracted deep features of I_m and I_f are flattened to feature tokens. $Adapter_0$ is the first to project the C dimensional features tokens in the visual domain to the language domain. It consists of two linear projections that gradually project the flattened feature tokens to the 4096 dimensions that the LLaMA block requires. The other $Adapter_i$ ($i \in 1, 2, 3, 4$) employs single linear projection to project the encoded feature tokens in the language domain to the visual domain. Each $Adapter_i$ ($i \in 1, 2, 3, 4$) project C dimensional features tokens from LLaMA block to $2^{2(i+1)}C$ dimensions.

3.3. LLaMA Block

LLaMA 2 is trained on the large-scale corpus dataset, which is an open-access LLM based on the transformer. According to the number of LLaMA 2 parameters, its pretrained weights can be divided into 7B, 13B, etc. We select the 7B pretrained weight for our work. Methods for fine-tuning language models include adapter [24], LoRA[42], etc. Adopting these technologies, the pretrained model can be applied to downstream tasks with only a small increase in the number of parameters. Inspired by [23], we use adapters to embed LLaMA 2 into our registration model. Then, we design a novel LLaMA block for our registration approach. As shown in Fig. 2, the LLaMA block consists of two pretrained LLaMA 2 that are connected by an *Inner Adapter*.

Inner Adapter is a Multi-Layer Perceptron (MLP), a module composed of multi-layer fully connected operations that scales features in the hidden layer. The feature dimensions of the pretrained LLaMA 2 7B input and output are both 4096. The previous LLaMA 2 encodes the projected visual features, and the output features are transformed into the hidden states using the *Inner Adapter*, then fed to the successive block. Therefore, by combining with the initial \mathcal{A}_{0} , the features in the visual domain can be converted into features in the language domain, and thus, the pretrained LLaMA2 can be applied for encoding. This projection and encoding can be formulated as

$$\mathcal{A}_0(x) \rightarrow x', \mathcal{F}_L^1(x') \rightarrow x', \mathcal{A}_{IN}(x') \rightarrow x', \mathcal{F}_L^2(x') \rightarrow y, \quad (2)$$

where x is the concatenated deep features in visual domain of I_m and I_f , \mathcal{A}_0 is Adapter₀ shown in Fig. 2, \mathcal{A}_{IN} is the *Inner Adapter*, and \mathcal{F}_L is the pretrained LLaMA 2. According to the process of Eq. 2, we can finally obtain feature y in the language domain modeled by the LLaMA block. It is worth noting that before this block, a learnable position embedding is added to make the LLaMA block compute the visual features with its position information.

3.4. Registration Model with Pretrained LLaMA 2

The proposed LLaMA-Reg and the proposed LLaMA block are shown in Fig. 2. In this work, we follow [43, 35] to construct a dual-stream encoder to extract the deep features separately, which captures the semantic correspondence of two volumes independently. In the dual encoding stage, the pyramid dual encoder extracts features of I_m and I_f . Specifically, each branch of the dual encoder has the same architecture. Both convolution blocks have a kernel size of 3. The convolution blocks with different strides of 1 or 2 are

for the same resolution feature extraction or downsampling. These features are represented by $F_m^i \in \{F_m^1, F_m^2, \dots, F_m^L\}$ and $F_f^i \in \{F_f^1, F_f^2, \dots, F_f^L\}$, where L represents the number of the resolution levels in the dual encoding stage.

3.4.1. Multi-Scale Registration Framework

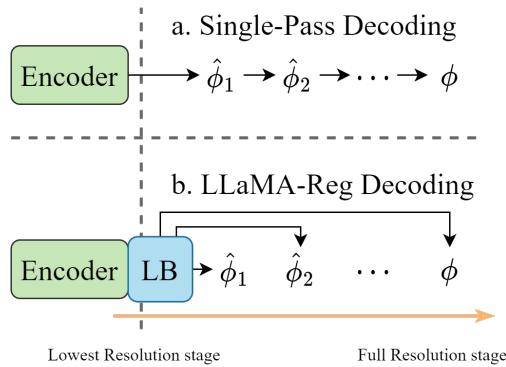


Figure 3: Two multi-scale decoding manners of Single-pass. a. is the normal decoding method, b. is the decoding method of LLaMA-Reg. LB is the LLaMA block.

The multi-scale registration framework predicts the deformation field separately from low to high resolution in a coarse-to-fine manner. The final deformation field is predicted using the composition of deformation fields at different scales. The deformation field in the high-resolution stage is based on the deformation field predicted in the coarse low-resolution stage, and the more refined deformation field is predicted. Therefore, the multi-scale registration framework can predict a larger displacement field than the single-stage registration model [14]. The single-pass framework [35] inspired us to design a novel multi-scale registration model, LLaMA-Reg. Unlike it, as shown in Fig. 3, our registration model leverages the modeling capabili-

ties of LLaMA to model the deep features of image pairs. At the same time, adapters are used to map features to different resolution stages and then generate deformation fields to perform multi-scale registration in the decoder.

We define S_4, S_3, S_2 and S_1 indicate the 1/8, 1/4, 1/2 and full resolution stages, respectively. To ensure the dimensions of the output features match the number of dimensions of each resolution stage, the dimensions output by the adapter in each decoder remain consistent with the number of features of the corresponding resolution stage after reorganization. As shown in Fig. 2, each stage utilizes the features encoded by the LLaMA block. First, stage S_i obtains the encoded feature tokens in the language domain, and the $Adapter_i$ projects them into $2^{2(i+1)}C$ dimensions, where $i \in \{1, 2, 3, 4\}$. The projected feature tokens in S_i are then reconstructed to the shape of previous resolution stages, with the number of dimensions consistent with $2^{2(i+1)}C$. Second, in S_i , the reconstructed features are upsampled to the shape of the next resolution stage using transposed convolution operations with a kernel size and stride of 2. After upsampling, features are concatenated with the corresponding features of both I_m and I_f from the dual encoding stage. The concatenated features are computed and fused by a convolution block, followed by another convolution block to generate the deformation fields in S_i . Both convolution blocks have a kernel size of 3 and a stride of 1 to maintain the feature shape. Furthermore, the upsampling operation scales the fused feature to the next stage shape. It is worth noting that in S_3 and S_4 , the deformation field and warped image obtained by composition in the previous stage are added to the feature fusion. In summary, based on the computation mentioned above,

the calculation of the fusion of the features and previous information in each S_i can be formulated as follows:

$$S_1 : F^4 = \text{Conv}(F_m^4, F_f^4, F_t^4), \hat{F}^{4'} = \text{Up}(F^4), \quad (3)$$

$$S_2 : F^3 = \text{Conv}(F_m^3, F_f^3, \hat{F}^{4'}, F_t^3), \hat{F}^{3'} = \text{Up}(F^3), \quad (4)$$

$$S_3 : F^2 = \text{Conv}(\hat{\phi}_1, \hat{I}_m^1, F_m^2, F_f^2, \hat{F}^{3'}, F_t^2), \hat{F}^{2'} = \text{Up}(F^2), \quad (5)$$

$$S_4 : F^1 = \text{Conv}(\hat{\phi}_2, \hat{I}_m^2, F_m^1, F_f^1, \hat{F}^{2'}, F_t^1) \quad (6)$$

In the multi-scale registration branch, deformation fields are represented by $\hat{\phi}_i$. The warped moving image at resolution stage S_i can be obtained by $\hat{I}_m^{i-1} = I_m^{i-1} \circ \hat{\phi}_{i-1}$ starting from S_2 , where I_m^{i-1} is the downsampled image at the corresponding stage, and \circ is the warping function (i.e., STN). At S_4 , the full resolution stage, ϕ is computed by a convolution block, and the final warped image is $\hat{I}_m = I_m \circ \hat{\phi}$.

3.4.2. Cascaded Registration Decoder

The first model in our cascaded approach is shown in Fig. 2. To further enhance the registration performance, we design a novel cascaded framework. Different from normal cascaded image registration approaches, our cascaded LLaMA-Reg consists of the LLaMA-Reg and the cascaded decoding branches, where branches share the same dual encoding branch. The cascaded framework of our LLaMA-Reg is shown in Fig. 4. This cascaded framework splits the registration procedure into n steps, and n decoders indicate n steps of the cascaded framework. These n steps branches share the same pretrained LLaMA block and trained encoder in *Step 1*. To train LLaMA-Reg efficiency and boost performance, the adapters, including *inner adapters*, and decoders are unfrozen in a cascaded training phase. Adapter_n^i

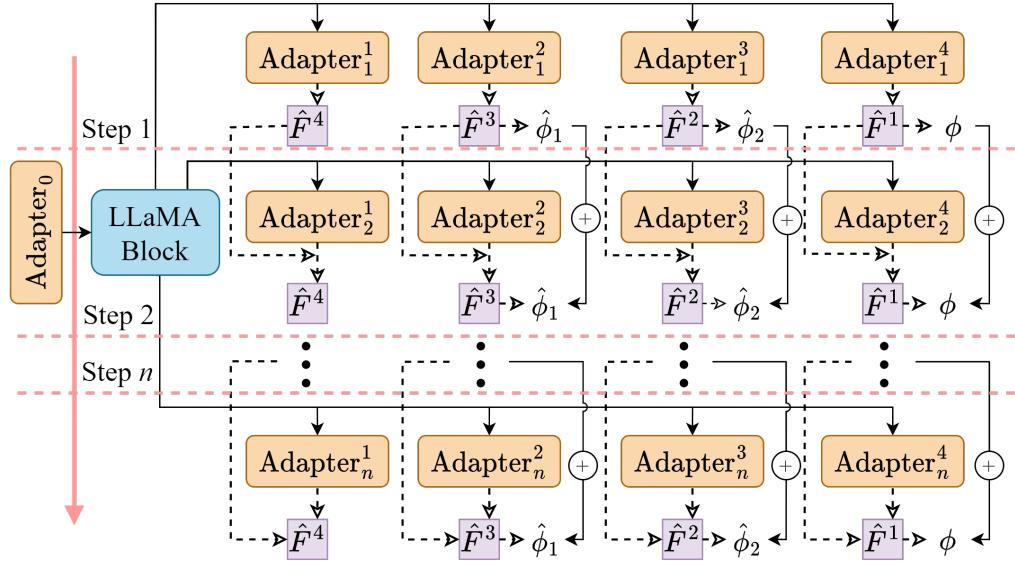


Figure 4: Overview of our proposed cascaded LLaMA-Reg. The registration procedure is split into n steps. The dotted arrows indicate the omission of some intermediate calculation processes.

is the adapter at i th resolution stage of step n . The feature used to generate the deformation field at each resolution stage is represented as $\hat{F}_n^{(5-i)}$. In the cascaded method we propose, the feature $\hat{F}_n^{(5-i)}$ generated in the previous step n participates in the feature fusion process of the $n+1$ step. Based on this, the feature $\hat{F}_{n+1}^{(5-i)}$ generated in the next step can be obtained. Besides the feature fusion between features at different steps, we also added the fusion of deformation fields at different steps. Specifically, the deformation field predicted in the previous step directly affects the deformation field generated in the next step. Thus, we can obtain the final deformation field in the last step by fusing these features and superposing the deformation field.

4. Experiments

4.1. Dataset Preparation

To validate the proposed method, LLaMA-Reg, we performed subject-to-subject registration tasks using two publicly available datasets: the OAI dataset ² for knee MR image registration and the OASIS dataset [44] for brain MR image registration.

Knee Subject-to-Subject Registration. For the knee subject-to-subject registration case, we employed the OAI dataset. This data repository provides images from an eleven-year longitudinal cohort study of knee osteoarthritis. Corresponding segmentation maps of knee MR images, obtained from [45], were utilized to estimate registration performance. The segmentation maps include femoral bone, tibial bone, femoral cartilage, and tibial cartilage. ANTs [46] was used to apply affine transformation to align each scan and then resample them to a size of $160 \times 160 \times 96$. The first 200 MR scans were used as the training set, and 90 MR scans were randomly selected from the remaining data (40 scans as fixed and 50 as moving images) to form a test set with a total of 2000 image pairs. For the test set, we invited a professional doctor to annotate the patella for each scan. Thus, the training set consists of 39000 image pairs, and the test set consists of 2000 image pairs, both with five labels, to train and evaluate registration performance.

Brain Subject-to-Subject Registration. To further demonstrate the performance of LLaMA-Reg, we conducted brain subject-to-subject registration on the OASIS [44] dataset, which is widely used in deep learning-based regis-

²<https://nda.nih.gov/oai/>

tration research. The OASIS data was obtained from Learn2Reg [47]. Each image has 35 anatomical segmentation maps in OASIS. We resampled the MRI scans into the shape of $112 \times 160 \times 128$. We followed the partitioning of OASIS in Learn2Reg to train the baseline and our models. Since the test set of OASIS in Learn2Reg is not public, we used the validation set to measure the subject-to-subject registration performance. A total of 380 image pairs can be generated through the validation set.

4.2. Baseline Methods

We compared LLaMA-Reg with four deep learning-based methods: Vox-
elMorph [31], LapIRN [14], TransMorph [10], and TransMatch [39].

VoxelMorph is a CNN-based registration model. It was the first to use a U-shaped structure to predict the full-resolution deformation field and applies the STN to warp the image through the deformation field to achieve image registration.

LapIRN is a coarse-to-fine registration model that learns multi-scale de-
formations. It is divided into three sub-models, predicting the deformation
field from $1/4$ to full resolution stages. Each sub-model is trained based on
the previous model to achieve a coarse-to-fine registration process, generating
large deformations.

TransMorph employs transformers to capture long-range correspondences
between voxels of an image pair. It is a hybrid model consisting of CNN
and Swin transformer components, demonstrating superior performance over
single-scale registration methods based on CNNs.

TransMatch utilizes a dual-stream framework to encode moving and fixed
images and includes a local window cross-attention module to achieve explicit

feature matching. This design leverages image feature matching information to enhance inter-image matching, further improving image registration. TransMatch also uses a multi-scale registration framework.

All baseline methods were trained for 300K iterations on the knee dataset and 400K iterations on the brain dataset. Using the MSE loss function, we set the coefficient to 0.04 for the OAI dataset and 0.02 for the OASIS dataset to better fit each dataset.

4.3. Implementation Details

We conducted experiments on a computing platform of Ubuntu server 23 with an NVIDIA RTX 3090 GPU. Our code was implemented using PyTorch 2.0. All baseline methods were trained using the Adam optimizer, with the learning rate set to 0.0001. Our method applied the similarity loss function MSE and the spatial gradient loss regularization term to train. When the hyperparameter of the spatial regularization term was set to 0.04 and 0.02 for knee and brain image registration, LLaMA-Reg performed better. The cascaded steps of LLaMA-Reg were set to 3. LLaMA-Reg was trained in 130K, 130K, and 30K iterations on the knee dataset for each cascaded step, respectively. For the brain image registration task, each step was set to 40K, 20K, and 20K, respectively. The pretrained LLaMA 2 was obtained from the Meta website³, as the LLaMA-Reg deep feature encoder, which requires the number of input and output dimensions are both 4096. It is worth noting that our model and the ablations were trained using PyTorch’s automatic mixed precision strategy. Our code is publicly available at [GitHub](#).

³<https://ai.meta.com/resources/models-and-libraries/>

4.4. Evaluation Metrics

We utilized the Dice metric to measure the registration accuracy of the proposed and baseline models. The Dice metric is calculated using the overlap of the corresponding segmentation labels of two images, with values ranging from 0 to 1. A higher Dice score indicates better registration accuracy. The non-positive Jacobian determinant is used to measure folds in a deformation. If the Jacobian determinant of the deformation field is non-positive, it indicates that the area will be folded during deformation. Time consumption is calculated as the inference time required to perform registration on the GPU for each pair of images.

4.5. Experiments

4.5.1. Experimental Results on OAI

We replicated the baseline methods and retrained all methods using the same dataset partition, then compared their performance on the same test set. The quantitative experimental results are shown in Table ???. Among these methods, the registration accuracy of the early CNN-based single-scale registration model, VoxelMorph, was lower than other recent registration methods. Among these methods, the registration accuracy of the early CNN-based single-scale registration model, VoxelMorph, was lower than that of more recent registration methods. Although LapIRN is also based on CNN, its accuracy was better than VoxelMorph and TransMoprh due to its multi-scale structure, which can effectively predict larger deformations. Using the powerful modeling capabilities of Transformers, TransMorph and TransMatch could construct distant voxel relationships, resulting in better registration performance than VoxelMorph. TransMatch introduced a

Table 1: Comparison results of subject-to-subject knee image registration task. Higher Dice (%) results indicate higher registration accuracy. $|J_\phi| \leq 0$ (%) indicates the percentage of folding voxels in a deformation. Time represents the registration seconds consumption using a trained method. Affine Only is the initial alignment result without registration.

Method	Knee			Brain	
	Dice	$ J_\phi \leq 0$	Time	Dice	$ J_\phi \leq 0$
Affine Only	32.10	—	—	49.92	—
VoxelMorph	61.63	0.22	35.19	76.42	0.17
TransMorph	64.93	0.29	65.63	77.53	0.17
LapIRN	65.26	0.10	46.79	77.39	0.10
TransMatch	66.02	0.19	82.03	76.67	0.19
LLaMA-Reg-C1 (Ours)	68.98	0.17	138.81	77.80	0.18
LLaMA-Reg (Ours)	71.55	0.26	241.53	78.40	0.17

multi-layer feature matching method in the transformer, which allowed it to capture deformation information at different feature levels, thereby predicting a more accurate deformation field than other baseline methods. Our method, LLaMA-Reg-C1, which is the proposed model without cascaded decoders, achieved better registration accuracy on the knee dataset. Compared with baseline methods on the Dice metric, LLaMA-Reg-C1 outperformed TransMatch by 2.96%, LapIRN by 3.72%, TransMorph by 4.05%, and VoxelMorph by 7.35%. Compared with TransMatch, TransMorph, LapIRN, and VoxelMorph, the proposed LLaMA-Reg, which has three cascaded decoders, showed performance improvements of 5.53%, 6.29%, 6.62%, and 9.92%, respectively. By applying the cascaded decoders, which are based on multiple adapters, LLaMA-Reg showed a significant improvement, with the Dice

metric increasing from 68.98% to 71.55% for knee scans and from 77.80% to 78.40% for brain scans.

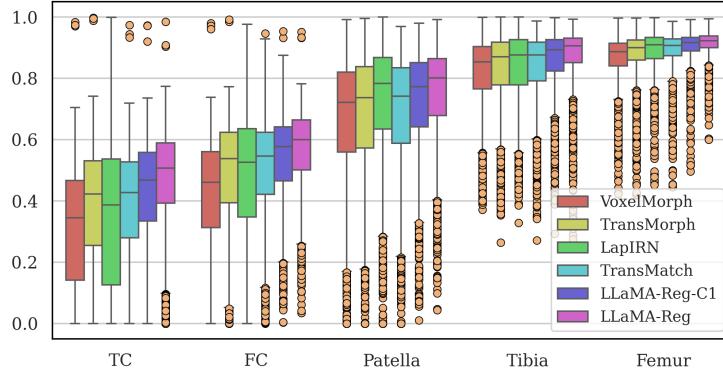


Figure 5: Box plot of registration results for all methods on each label. TC and FC represent tibia cartilage and femur cartilage, respectively.

We presented the statistical results of all labels in Fig. 5. Except for the Dice results of LapIRN on the patella, which were higher than our method, ours both achieved superior registration accuracy. The experimental results on the OAI dataset showed that using the multi-scale model combined with adapters and pretrained LLaMA 2 to map the features of an image pair into the language domain could more accurately predict a deformation field. The results of $|J_\phi| \leq 0(\%)$ indicated that the results of all methods are on the same order of magnitude, but the results of LapIRN were smoother. In the column of time consumption, we found that CNN-based methods had less inference time consumption than ViT-based methods. Since the more complex architectures of our methods, there was an increase in inference time.

We presented the statistical results for all labels in Fig. 5. Except for the Dice results of LapIRN on the patella, which were higher than our method,

our methods both achieved superior registration accuracy. The experimental results on the OAI dataset showed that using the multi-scale model combined with adapters and the pretrained LLaMA 2 to map the features of an image pair into the language domain could more accurately predict a deformation field. The results of $|J_\phi| \leq 0(\%)$ indicated that the results of all methods are of the same order of magnitude, but the results of LapIRN were smoother. In terms of time consumption, we found that CNN-based methods had less inference time than ViT-based methods. Due to the more complex architectures of our methods, there was an increase in inference time.

The quantitative results are shown in Fig. 6. From the color bars of deformations, we noted that our method predicted the largest displacement among these methods, indicating that our method established the correspondence between more distant voxels. Through the warped images, it can be seen that among all methods, our method produced the most similar warped slice to the fixed image, especially the comparison of the bone structure represented by the labels.

4.5.2. Additional Experimental Results on OASIS

We additionally tested our methods on the OASIS dataset. As shown in Table 1, our methods also achieved the best registration accuracy. In this experiment, the result of LapIRN on the $|J_\phi| \leq 0(\%)$ metric was optimal, and the results of other methods were similar. The result of LLaMA-Reg-C1 was better than VoxelMorph, TransMorph, LapIRN, and TransMatch by 1.38%, 0.27%, 0.41%, and 1.13%, respectively. The registration performance of the cascaded LLaMA-Reg further improved on the brain dataset. Compared with these methods, the performance improved by 1.98%, 0.97%, 1.01%, and

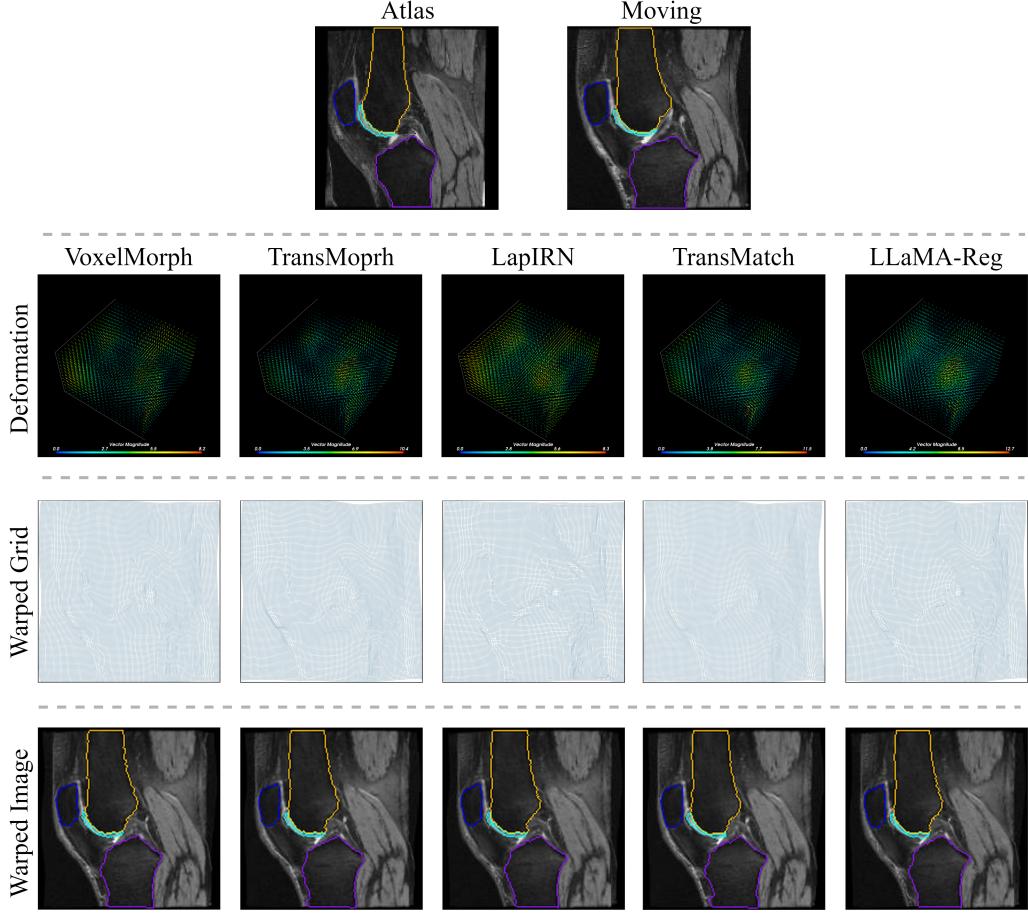


Figure 6: Visualization of experimental results on knee MRI. Deformation represents the displacement direction and magnitude of pixels in 3D MR; the warped grid reflects the changes in the current slice. In the warped images, the patella, femur, femoral cartilage, and tibia are represented in blue, yellow, light blue, and purple, respectively.

1.73%, respectively. By applying cascaded decoders, LLaMA-Reg had an enhancement of 0.6%. We averaged the Dice scores of symmetric structures in OASIS, combining the Dice metrics of 36 labels into 19. The box plot of the statistical results on the brain is shown in Fig. 7. These results illustrated that our methods performed best on most segmentation maps.

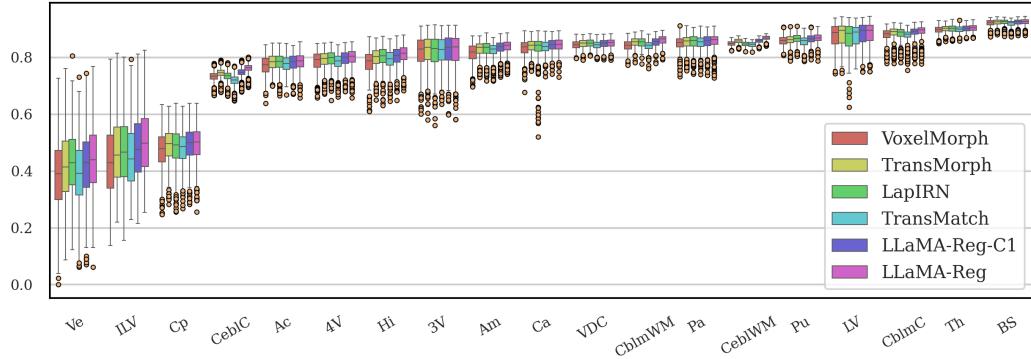


Figure 7: Box plot of registration results for all methods on OASIS.

4.6. Ablation Studies

To verify the impact of the various proposed schemes on registration performance and their effectiveness from multiple aspects, we conducted several ablation experiments of LLaMA-Reg-C1 utilizing the OAI dataset. Ablation results are shown in 2.

Table 2: Three ablation studies of LLaMA-Reg-C1. Our ablation experiments report the registration accuracy and the GPU memory occupation during training.

	Model	Dice	Memory (MB)
Ablation 1	LLaMA 2	67.39	10474
	Pretrained LLaMA 2	68.98	7632
	Standard ViT	67.51	22520
Ablation 2	w/o Pos. Emb	59.37	7208
	Dim (4096×4)	69.08	7750
	Dim (4096×2) & w/ Pos. Emb	68.98	7632
Ablation 3	Joint training	71.36	14712
	Step-by-step	71.55	8384 (3rd Decoder)

Verify the performance improvement of pretrained LLaMA 2

and model architecture. To demonstrate that pretrained LLaMA 2 can improve the modeling ability of the medical image registration model, we compared the registration performance of using LLaMA 2 Transformer blocks and pretrained LLaMA 2. The results of Ablation 1 indicated that the performance of the registration model using pretrained LLaMA 2 was better than using the LLaMA 2 Transformer blocks when extracting deep features. Additionally, using pretrained models can reduce memory usage and speed up training.

Furthermore, to investigate whether the scheme of our architecture had a performance advantage, we replaced LLaMA 2 blocks with standard ViT blocks. The configuration of ViT blocks used for replacement were: number of channels 4096, number of heads 4, number of depth 2. The registration accuracy in Table 1, showed that compared with the scheme using LLaMA Transformer blocks and pretrained LLaMA 2, the performance declined, but it still outperformed baseline methods. This demonstrated that our registration framework with a non-U-shaped structure could achieve higher registration accuracy, proving that the proposed non-U-shaped architecture is more suitable for registration tasks.

Imapct of Inner settings for Performance When we tested the impact of the MLP mapping scale on the modeling performance of LLaMA 2, we found that although expanding the mapping scale can improve performance, the degree of improvement was limited. Therefore, we chose a mapping multiple of 2 in this work.

In the standard ViT, when using linear projection to model image features, the spatial information of these features should be modeled simultane-

ously. Therefore, when using transformers to solve visual problems, learnable positional encodings are added to express the spatial information of image tokens. We believe that positional embeddings need to be added when calculating image features because adapters, which consist of linear projections, need to learn spatial information. In Ablation 2, we removed the standard positional embedding in LLaMA-Reg-C1. It was observed that without positional embedding, the accuracy of registration dropped significantly. This shows that when using pretrained LLaMA 2 to calculate visual features, positional encoding is also necessary to record the positional information of the image.

Step-by-step training and joint training for the cascaded decoder. We trained the cascaded decoder in two ways: step-by-step training and joint training. In step-by-step training, the previously trained network is frozen while training the next network. For joint training, the weights of the dual encoding branches were frozen, and all decoding branches were trained together. The different results of the two training manners are shown in Ablation 3. We reported the maximum GPU memory usage during step-by-step training (i.e., the 3rd cascaded step). When the number of cascaded decoding steps was set to 3, the step-by-step trained model outperformed the jointly trained model. The step-by-step trained model only needed to train the decoding part of the current step in each training stage, which occupied less GPU memory and generated more accurate deformations.

5. Conclusion

In this work, we propose to use the large language model LLaMA2 as the deep feature calculation component of the registration model and propose an unsupervised medical image registration model with a non-U-shaped structure. In order to use the features calculated by LLaMA2, an adapter is used to convert visual features and language features into each other in order to transfer the calculated features to each scale stage for multi-scale registration. Experimental results on knee and brain data show that our method achieves optimal results. Through ablation experiments, we demonstrate the effectiveness of our proposed non-U-shaped registration framework and the use of pre-trained LLaMA2.

6. Acknowledgements

We would like to express our special gratitude to Dr. Yu Yang from the Department of Orthopedics at Taizhou Hospital for his invaluable contributions to this work, particularly in areas involving medical knowledge and medical imaging. Dr. Yang professionally annotated the test sets used for numerous experiments in the knee MRI registration task, enabling us to test all methods rigorously. Additionally, understanding the changes in bones and joints in knee imaging is crucial for analyzing and diagnosing diseases related to bones and joints. This work lays the foundation for our future research on diseases involving morphological changes in knee bones. This study was funded by the Zhejiang Province Medical Science and Technology Program of China (No. 2020PY088), the Scientific Research of Enze Medical Center (Group) (No. 24EZA01), the Scientific Research of Enze Medical

Center (Group) (No. 24EZJX02), the Scientific Research of Enze Medical Center (Group) (No. 24EZCG03)

References

- [1] N. Decaux, P.-H. Conze, J. Ropars, X. He, F. T. Sheehan, C. Pons, D. B. Salem, S. Brochard, F. Rousseau, Semi-automatic muscle segmentation in mr images using deep registration-based label propagation, *Pattern Recognition* 140 (2023) 109529.
- [2] P. Alvarez, S. Rouzé, M. I. Miga, Y. Payan, J.-L. Dillenseger, M. Chabanas, A hybrid, image-based and biomechanics-based registration approach to markerless intraoperative nodule localization during video-assisted thoracoscopic surgery, *Medical image analysis* 69 (2021) 101983.
- [3] Y.-M. Lin, I. Paolucci, C. S. O'Connor, B. M. Anderson, B. Rigaud, B. M. Fellman, K. A. Jones, K. K. Brock, B. C. Odisio, Ablative margins of colorectal liver metastases using deformable ct image registration and autosegmentation, *Radiology* 307 (2023) e221373.
- [4] W. Nie, R. Chang, M. Ren, Y. Su, A. Liu, I-gcn: Incremental graph convolution network for conversation emotion detection, *IEEE Transactions on Multimedia* 24 (2022) 4471–4481. doi:[10.1109/TMM.2021.3118881](https://doi.org/10.1109/TMM.2021.3118881).
- [5] W. Nie, Y. Zhao, D. Song, Y. Gao, Dan: Deep-attention network for 3d shape recognition, *IEEE Transactions on Image Processing* 30 (2021) 4371–4383. doi:[10.1109/TIP.2021.3071687](https://doi.org/10.1109/TIP.2021.3071687).
- [6] F. Isensee, P. F. Jaeger, S. A. Kohl, J. Petersen, K. H. Maier-Hein, nnunet: a self-configuring method for deep learning-based biomedical image segmentation, *Nature methods* 18 (2021) 203–211.

- [7] J. M. Noothout, N. Lessmann, M. C. Van Eede, L. D. van Harten, E. Sogancioglu, F. G. Heslinga, M. Veta, B. van Ginneken, I. Işgum, Knowledge distillation with ensembles of convolutional neural networks for medical image segmentation, *Journal of Medical Imaging* 9 (2022) 052407–052407.
- [8] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in: N. Navab, J. Hornegger, W. M. Wells, A. F. Frangi (Eds.), *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, Springer International Publishing, Cham, 2015, pp. 234–241.
- [9] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, J. Liang, Unet++: Re-designing skip connections to exploit multiscale features in image segmentation, *IEEE transactions on medical imaging* 39 (2019) 1856–1867.
- [10] 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.
- [11] M. Ma, Y. Xu, L. Song, G. Liu, Symmetric transformer-based network for unsupervised image registration, *Knowledge-Based Systems* (2022) 109959.
- [12] 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, in: *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, 2022, pp. 217–226.
- [13] H. Sokooti, B. De Vos, F. Berendsen, B. P. Lelieveldt, I. Işgum, M. Starling, Nonrigid image registration using multi-scale 3d convolutional

- neural networks, in: 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. 232–239.
- [14] T. C. Mok, A. C. Chung, Large deformation diffeomorphic image registration with laplacian pyramid networks, in: Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part III 23, Springer, 2020, pp. 211–221.
 - [15] S. Zhao, Y. Dong, E. I.-C. Chang, Y. Xu, Recursive cascaded networks for unsupervised medical image registration, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019.
 - [16] S. Zhao, T. Lau, J. Luo, I. Eric, C. Chang, Y. Xu, Unsupervised 3d end-to-end medical image registration with volume tweening network, IEEE journal of biomedical and health informatics 24 (2019) 1394–1404.
 - [17] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, et al., Llama 2: Open foundation and fine-tuned chat models, arXiv preprint arXiv:2307.09288 (2023).
 - [18] S. Biderman, H. Schoelkopf, Q. G. Anthony, H. Bradley, K. O’Brien, E. Hallahan, M. A. Khan, S. Purohit, U. S. Prashanth, E. Raff, et al., Pythia: A suite for analyzing large language models across training and scaling, in: International Conference on Machine Learning, PMLR, 2023, pp. 2397–2430.

- [19] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., Language models are few-shot learners, *Advances in neural information processing systems* 33 (2020) 1877–1901.
- [20] J. Li, D. Li, S. Savarese, S. Hoi, Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, arXiv preprint arXiv:2301.12597 (2023).
- [21] K. Zhou, J. Yang, C. C. Loy, Z. Liu, Learning to prompt for vision-language models, *International Journal of Computer Vision* 130 (2022) 2337–2348.
- [22] K. Zhou, J. Yang, C. C. Loy, Z. Liu, Learning to prompt for vision-language models, *International Journal of Computer Vision* 130 (2022) 2337–2348.
- [23] S.-A. Rebuffi, H. Bilen, A. Vedaldi, Learning multiple visual domains with residual adapters, *Advances in neural information processing systems* 30 (2017).
- [24] N. Houlsby, A. Giurgiu, S. Jastrzebski, B. Morrone, Q. De Laroussilhe, A. Gesmundo, M. Attariyan, S. Gelly, Parameter-efficient transfer learning for nlp, in: International Conference on Machine Learning, PMLR, 2019, pp. 2790–2799.
- [25] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).
- [26] K. Zhou, J. Yang, C. C. Loy, Z. Liu, Learning to prompt for vision-language models, *International Journal of Computer Vision* 130 (2022)

2337–2348.

- [27] M. Trager, P. Perera, L. Zancato, A. Achille, P. Bhatia, S. Soatto, Linear spaces of meanings: compositional structures in vision-language models, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 15395–15404.
- [28] 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 (2008) 26–41.
- [29] M. P. Heinrich, M. Jenkinson, M. Brady, J. A. Schnabel, Mrf-based deformable registration and ventilation estimation of lung ct, *IEEE Transactions on Medical Imaging* 32 (2013) 1239–1248.
- [30] J. Duan, X. Jia, J. Bartlett, W. Lu, Z. Qiu, Arbitrary order total variation for deformable image registration, *Pattern Recognition* (2023) 109318.
- [31] 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 (2019) 1788–1800.
- [32] 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.
- [33] T. Mok, A. Chung, Fast symmetric diffeomorphic image registration with convolutional neural networks, in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020.
- [34] B. Kim, H. K. Dong, S. H. Park, J. Kim, J. C. Ye, Cyclemorph: Cycle

- consistent unsupervised deformable image registration, *Medical Image Analysis* 71 (2021) 102036.
- [35] M. Meng, L. Bi, D. Feng, J. Kim, Non-iterative coarse-to-fine registration based on single-pass deep cumulative learning, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2022, pp. 88–97.
 - [36] J. Chen, Y. He, E. Frey, Y. Li, Y. Du, Vit-v-net: Vision transformer for unsupervised volumetric medical image registration, in: *Medical Imaging with Deep Learning*, 2021.
 - [37] Y. Zhang, Y. Pei, H. Zha, Learning dual transformer network for diffeomorphic registration, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2021, pp. 129–138.
 - [38] Y. Zhu, S. Lu, Swin-voxelmorph: A symmetric unsupervised learning model for deformable medical image registration using swin transformer, in: L. Wang, Q. Dou, P. T. Fletcher, S. Speidel, S. Li (Eds.), *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*, Springer Nature Switzerland, Cham, 2022, pp. 78–87.
 - [39] Z. Chen, Y. Zheng, J. C. Gee, Transmatch: A transformer-based multi-level dual-stream feature matching network for unsupervised deformable image registration, *IEEE Transactions on Medical Imaging* (2023).
 - [40] M. Jaderberg, K. Simonyan, A. Zisserman, et al., Spatial transformer networks, *Advances in neural information processing systems* 28 (2015) 2017–2025.
 - [41] G. Balakrishnan, A. Zhao, M. R. Sabuncu, J. Guttag, A. V. Dalca,

An unsupervised learning model for deformable medical image registration, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018.

- [42] E. J. Hu, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, et al., Lora: Low-rank adaptation of large language models, in: International Conference on Learning Representations, 2021.
- [43] M. Kang, X. Hu, W. Huang, M. R. Scott, M. Reyes, Dual-stream pyramid registration network, Medical image analysis 78 (2022) 102379.
- [44] D. S. Marcus, T. H. Wang, J. Parker, J. G. Csernansky, J. C. Morris, R. L. Buckner, Open access series of imaging studies (oasis): cross-sectional mri data in young, middle aged, nondemented, and demented older adults, Journal of cognitive neuroscience 19 (2007) 1498–1507.
- [45] F. Ambellan, A. Tack, M. Ehlke, S. Zachow, Automated segmentation of knee bone and cartilage combining statistical shape knowledge and convolutional neural networks: Data from the osteoarthritis initiative, Medical image analysis 52 (2019) 109–118.
- [46] B. B. Avants, N. J. Tustison, G. Song, P. A. Cook, A. Klein, J. C. Gee, A reproducible evaluation of ants similarity metric performance in brain image registration, Neuroimage 54 (2011) 2033–2044.
- [47] A. Hering, L. Hansen, T. C. Mok, A. C. Chung, H. Siebert, S. Häger, A. Lange, S. Kuckertz, S. Heldmann, W. Shao, et al., Learn2reg: comprehensive multi-task medical image registration challenge, dataset and evaluation in the era of deep learning, IEEE Transactions on Medical Imaging 42 (2022) 697–712.