Affine Medical Image Registration with Coarse-to-Fine Vision Transformer

Tony C. W. Mok, Albert C. S. Chung
Department of Computer Science and Engineering,
The Hong Kong University of Science and Technology
cwmokab@connect.ust.hk, achung@cse.ust.hk

Abstract

Affine registration is indispensable in a comprehensive medical image registration pipeline. However, only a few studies focus on fast and robust affine registration algorithms. Most of these studies utilize convolutional neural networks (CNNs) to learn joint affine and non-parametric registration, while the standalone performance of the affine subnetwork is less explored. Moreover, existing CNN-based affine registration approaches focus either on the local misalignment or the global orientation and position of the input to predict the affine transformation matrix, which are sensitive to spatial initialization and exhibit limited generalizability apart from the training dataset. In this paper, we present a fast and robust learning-based algorithm, Coarse-to-Fine Vision Transformer (C2FViT), for 3D affine medical image registration. Our method naturally leverages the global connectivity and locality of the convolutional vision transformer and the multi-resolution strategy to learn the global affine registration. We evaluate our method on 3D brain atlas registration and template-matching normalization. Comprehensive results demonstrate that our method is superior to the existing CNNs-based affine registration methods in terms of registration accuracy, robustness and generalizability while preserving the runtime advantage of the learning-based methods. The source code is available at https://github.com/cwmok/C2FViT.

1. Introduction

Rigid and affine registration is crucial in a variety of medical imaging studies and has been a topic of active research for decades. In a comprehensive image registration framework, the target image pair is often pre-aligned based on a rigid or affine transformation before using deformable (non-rigid) registration, eliminating the possible linear and large spatial misalignment between the target image pair. Solid structures such as bones can be aligned well with rigid and affine registration [29, 37]. In conventional image registration approaches, inaccurate pre-alignment of the
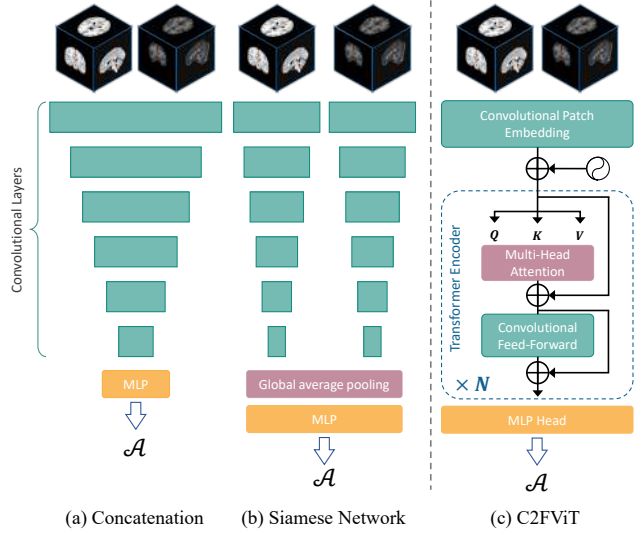

Figure 1. Comparisons of different architectures for affine registration. The concatenation-based (VTN-Affine [46]) and Siamese network (ConvNet-Affine [11]) approaches are based on convolutional neural networks, while our proposed C2FViT is based on vision transformers. For brevity, we illustrate 1-level C2FViT only. Local and global operations are in green and purple, respectively.

image pair may impair the registration accuracy or impede the convergence of the optimization algorithm, resulting in sub-optimal solutions [47]. The success of recent learning-based deformable image registration approaches has largely been fueled [3,9,11,17,19,20,34–36] by accurate affine initialization using conventional image registration methods. While the conventional approaches excel in registration performance, the registration time is dependent on the degree of misalignment between the input images and can be time-consuming with high-resolution 3D image volumes. To facilitate real-time automated image registration, a few studies [21, 22, 40, 46] have been proposed to learn joint affine and non-parametric registration with convolutional neural networks (CNNs). However, the standalone performance of

the affine subnetwork compared to the conventional affine registration algorithm is less explored. Moreover, considering that affine transformation is global and generally targets the possible large displacement, we argue that CNNs are not the ideal architecture to encode the orientation and absolution position of the image scans in Cartesian space or affine parameters due to the inductive biases embedded into the architectural structure of CNNs.

In this paper, we analyze and expose the generic inability and limited generlizability of CNN-based affine registration methods in cases with large initial misalignment and unseen image pairs apart from the training dataset. Motivated by the recent success of vision transformer models [10, 12, 41, 43, 44], we depart from the existing CNN-based approaches and propose a coarse-to-fine vision transformer (C2FViT) dedicated to 3D medical affine registration. To the best of our knowledge, this is the first learning-based affine registration approach that considers the non-local dependencies between input images when learning the global affine registration for 3D medical image registration.

The main contributions of this work are as follows:

- we quantitatively investigate and analyze the registration performance, robustness and generalizability of existing learning-based affine registration methods and conventional affine registration methods in 3D brain registration;

- we present a novel learning-based affine registration algorithm, namely C2FViT, which leverages convolutional vision transformers with the multi-resolution strategy. C2FViT outperforms the recent CNN-based affine registration approaches while demonstrating superior robustness and generalizability across datasets;

- the proposed learning paradigm and objective functions can be adapted to a variety of parametric registration approaches with minimum effort.

We evaluate our method on two tasks: template-matching normalization to MNI152 space [13–15] and 3D brain atlas registration in native space. Results demonstrate that our method not only achieves superior registration performance over existing CNN-based methods, but the trained model also generalizes well to an unseen dataset beyond the training dataset, reaching the registration performance of conventional affine registration methods.

2. Related Work

2.1. Learning-based Affine Registration Methods

Conventional approaches often formulate the affine registration problem to an iterative optimization problem, which optimizes the affine parameters directly using adaptive gradient descent [1, 25] or convex optimization [18].

While conventional approaches excel in registration accuracy, the registration time is subject to the complexity and resolution of the input image pairs. Recently, many learning-based approaches have been proposed for fast affine registration. These approaches significantly accelerate the registration time by formulating the affine registration problem as a learning problem using CNNs and circumventing the costly iterative optimization in conventional approaches. Existing CNN-based affine registration approaches can be divided into two categories: concatenation-based [21, 22, 33, 46] and Siamese network approaches [5, 11, 38] as shown in figure 1.

Zhao et al. [46] propose a concatenation-based affine subnetwork that concatenates the fixed and moving images as input, and exploits single-stream CNNs to extract the features based on the local misalignment of the input. Considering affine registration is global, their method is not capable of input with large initial misalignment as the affine subnetwork lacks global connectivity and only focuses on the overlapping region between two image spaces. In contrast to the concatenation-based method, de Vos et al. [11] propose an unsupervised affine registration method using the Siamese CNN architecture for fixed and moving images. A global average pooling [27] is applied to the end of each pipeline in order to extract one feature per feature map, forcing the networks to encode orientations and affine transformations globally. Although their network focuses on the global high-level geometrical features of separated input, their method completely ignores the local features of the initial misalignment between the input image pair. Moreover, a recent study [28] demonstrates that a pure CNN encoder fails spectacularly in a seemingly trivial coordinate transform problem, implying that a pure CNN encoder may not be an ideal architecture to encode the orientations and absolution positions of the image scans in Cartesian space or to affine parameters. Shen et al. [40] also report that CNN-based affine registration methods do not perform well in practice, even for deep CNNs with large receptive fields.

It is worth noting that most of the existing CNN-based affine registration methods [5, 11, 21, 22, 38, 46] jointly evaluate the affine and deformable registration performance or completely ignore the standalone performance of the affine subnetwork compared to the conventional affine registration algorithms. As inaccurate affine pre-alignment of the image pair may impair the registration accuracy or impede the convergence of the deformable registration algorithm [40, 47], a comprehensive evaluation of the CNN-based affine registration methods should by no means be ignored.

2.2. Vision Transformer

CNNs architecture generally has limitations in modelling explicit long-range dependencies due to the intrinsic inductive biases, *i.e.*, weight sharing and locality, embedded

into the architectural structure of CNNs. Recently, Dosovitskiy et al. [12] proposed a pioneering work, Vision Transformer (ViT), for image classification and proved that a pure transformer [41] architecture can attain a state-of-the-art performance. Compared to CNN-based approaches, ViT offers less image-specific inductive bias and has tremendous potential when training in large scale datasets. Wang et al. [43] develop a pyramid architectural design for a pure transformer model to imitate the multi-scale strategy in CNNs, achieving promising results in various computer vision tasks. Subsequent studies [6–8, 10, 16, 26, 42, 44] further extend ViT to pyramid architectural design and introduce convolutions to ViT. These studies demonstrate that introducing moderate convolutional inductive bias to ViT improves the overall performance, especially for training with small datasets. Apart from pure ViT methods, Zhang et al. [45] and Chen et al. [4] combine CNN encoder-decoder with transformer for deformable registration.

While CNNs have achieved remarkable success in deformable medical image registration, we argue that CNNs are not an ideal architecture for modelling and learning affine registration. In contrast to deformable image registration, affine registration is often used to mitigate and remove large linear misalignment, which is considered to be a global operation and contradicts the inductive bias embedded in the architectural structure of CNNs. Building on the insights of ViT and its variants [10, 12, 43, 44], we depart from the CNNs architecture and propose a pure transformer-based method dedicated to 3D medical affine registration.

3. Method

Let F, M be fixed and moving volumes defined over a n-D mutual spatial domain $\Omega \subseteq \mathbb{R}^n$. In this paper, we focus on 3D affine medical image registration, *i.e.*, $n = 3$ and $\Omega \subseteq \mathbb{R}^3$. For simplicity, we further assume that F and M are single-channel, grayscale images. Our goal is to learn the optimal affine matrix that align F and M. Specifically, we parametrized the affine registration problem as a function $f_\theta(F, M) = \mathcal{A}$ using a coarse-to-fine vision transformer (C2FViT), where θ is a set of learning parameters and $\mathcal{A}$ represents the predicted affine transformation matrix.

3.1. Coarse-to-fine Vision Transformer (C2FViT)

The overall pipeline of our method is depicted in figure 2. Our method has been divided into L stages that solves the affine registration in a coarse-to-fine manner with an image pyramid. All stages share an identical architecture consisting of a *convolutional patch embedding* layer and N_i transformer encoder blocks, where N_i denotes the number of transformer blocks in stage i. Each transformer encoder block consists of an alternating multi-head self-attention module and a *convolutional feed-forward layer*, as depicted

in figure 1. We use $L = 3$ and $N_i = 4$ for each stage i throughout this paper. Specifically, we first create the input pyramid by downsampling the input F and M with trilinear interpolation to obtain $F_i \in \{F_1, F_2, \ldots, F_L\}$ (and $M_i \in \{M_1, M_2, \ldots, M_L\}$), where F_i represents the downsampled F with a scale factor of 0.5^{L-i} and $F_L = F$. We then concatenate F_i and M_i, and the concatenated input is subjected to the convolutional patch embedding layer. Different from the prior Transformer-based architectures [10, 12, 43, 44], we prune all the layer normalization operations as we did not observe noticeable effects on the image registration performance in our experiments. Next, a stack of N_i transformer encoder blocks take as input the image patch embedding map and output the feature embedding of the input. C2FViT solves the affine registration problem in a coarse-to-fine manner, and the intermediate input moving image M_i is transformed via *progressive spatial transformation*. Additionally, for stage $i > 1$, a residual connection from the output embeddings (tokens) of the previous stage $i - 1$ is added to the patch embeddings of the current stage i. Finally, the estimated affine matrix $\mathcal{A}_L$ of the final stage is adopted as the output of our model f_θ.

3.1.1 Locality of C2FViT

While the ViT model [12] excels in modelling long-range dependencies within a sequence of non-overlapping image patches due to the self-attention mechanism, the vision transformer model lacks locality mechanisms to model the relationship between the input patch and its neighbours. Therefore, we follow [26, 42, 44] to add locality to our transformers in C2FViT. Specifically, we mainly improve the transformer in two aspects: patch embedding and feedforward layer.

As shown in figure 2, we depart from the linear patch embedding approach [12] and adopt convolutional patch embedding [42, 44] instead. The goal of the convolutional patch embedding layer is to convert the input images into a sequence of overlapping patch embeddings. Formally, given a concatenated input $I \in \mathbb{R}^{H \times W \times D \times C}$, where H, W and D denote the spatial dimension of I, and C is the number of channels, the convolutional patch embedding layer utilizes a 3D convolution layer to compute the patch embedding map $\mathbf{Z} \in \mathbb{R}^{H_i \times W_i \times D_i \times d}$ of I. Specifically, the kernel size, stride, number of zero-paddings and number of feature maps of the 3D convolution layer are denoted as k^3, s, p and d, respectively. Next, the patch embedding map $\mathbf{Z}$ is then flattened into a sequence of patch embeddings (tokens) $\{\hat{\mathbf{Z}}_i \in \mathbb{R}^d | i = 1, \ldots, N\}$, where $N = H_i W_i D_i$ and d is the embedding dimension. The patch embeddings can be aggregated into a matrix $\hat{\mathbf{Z}} \in \mathbb{R}^{N \times d}$. We restrict the number of patches N to 4096 and the embedding dimension d to 256 for all convolutional patch embedding layers in C2FViT by varying the stride s of the convolution layer,
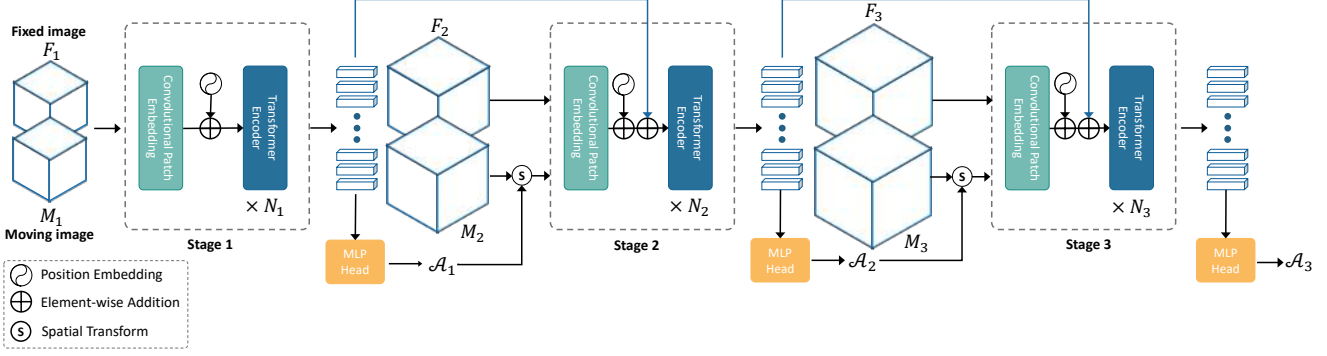
Figure 2. Overview of the proposed Coarse-to-Fine Vision Transformer (C2FViT). The entire model is divided into three stages, solving the affine registration in a coarse-to-fine manner.

i.e., $s = (\frac{H}{16}, \frac{W}{16}, \frac{D}{16})$. Moreover, we enforce the window overlapping to the sliding window of the convolution operation by setting k to $2s - 1$, and pad the feature with zeros ($p = \lfloor \frac{k}{2} \rfloor$). In contrast to the linear patch embedding in ViT, the convolutional patch embedding in C2FViT helps model local spatial context and features across the fixed and moving images. It also provides flexibility to adjust the number and feature dimensions of patch embeddings. On the other hand, the feed-forward layer in ViT consists of a MLP block with two hidden layers. In the transformer encoder, the feed-forward layer is the only local and translation equivariance. Since the feed-forward layer in ViT is applied to the patch embeddings map in a patch-wise manner, it lacks a local mechanism to model the relationship between adjacent patch embeddings. As such, we add a $3 \times 3 \times 3$ depthwise convolution layer in between two hidden layers of a MLP block in the feed-forward layer of C2FViT [26, 42]. The depth-wise convolution further introduces locality into the transformer encoder of C2FViT.

3.1.2 Global Connectivity of C2FViT

Transformers excel in modelling long-range dependencies within a sequence of embedding owing to their self-attention mechanism. In contrast to existing CNN-based affine registration approaches, the misalignment and the global relationship between the fixed and moving images can be captured and modelled by the similarity between the projected query-key pairs in transformer encoders of C2FViT, yielding the attention score for each patch embedding. Specifically, the query $\mathbf{Q}$, key $\mathbf{K}$, and value $\mathbf{V}$ are a linearly projection of the patch embeddings (tokens), *i.e.*, $\mathbf{Q} = \hat{\mathbf{Z}}\mathbf{W}^Q$, $\mathbf{K} = \hat{\mathbf{Z}}\mathbf{W}^K$ and $\mathbf{V} = \hat{\mathbf{Z}}\mathbf{W}^V$. We further extend the self-attention module to a multi-head self-attention (MHA) module [41]. Given the number of attention heads is h, the linear projection matrices $\mathbf{W}_j^Q$, $\mathbf{W}_j^K$ and $\mathbf{W}_j^V$ for each attention head j are the same size, *i.e.*, $\mathbf{W}_j^Q$, $\mathbf{W}_j^K$, $\mathbf{W}_j^V \in \mathbb{R}^{d \times d_h}$ and $d_h = \frac{d}{h}$. Following the self-attention

mechanism [12, 41] in the original transformer, our attention operation for attention head j is computed as:

$$\text{Attention}(\mathbf{Q}_j, \mathbf{K}_j, \mathbf{V}_j) = \text{Softmax}(\frac{\mathbf{Q}_j\mathbf{K}_j^\mathsf{T}}{\sqrt{d_h}})\mathbf{V}_j \quad (1)$$

where d_h is the embedding dimension for the attention head. At the end, the attended embeddings of all attention heads are concatenated and linear projected by a matrix $\mathbf{W}^O \in \mathbb{R}^{d \times d}$. In this study, we employ $h = 2$ attention heads and $d = 256$ embedding dimension for all the transformer encoders.

3.1.3 Progressive Spatial Transformation

We adopt the multiresolution strategy into our architectural design. Specifically, a classification head, which is implemented by two successive multilayer perceptrons (MLP) layers with the hyperbolic tangent (Tanh) activation function, is appended at the end of each stage in C2FViT. The classification head takes as input the averaged patch-wise patch embedding and outputs a set of affine transformation parameters. In the intermediate stage i, the derived affine matrix is used to progressively transform the moving image M_{i+1} with a spatial transformer [23]. The warped moving image M_{i+1} is then concatenated with fixed image F_{i+1} and taken as input for stage $i + 1$. With the proposed progressive spatial transformation, the linear misalignment of the input images can easily be eliminated with low-resolution input, and the transformers from the higher level can focus on the complex misalignment between the input image pair, reducing the complexity of the problem at the higher stages.

3.2. Decoupled Affine Transformation

While directly estimating the affine matrix is feasible [21, 38, 46], this transformation model cannot generalize

to other parametric registration methods as the affine matrix cannot decompose into a set of linear geometric transformation matrices, *i.e.*, translation, rotation, scaling and shearing. In the transformation model of C2FViT, we take a step further and utilize C2FViT to predict a set of geometric transformation parameters instead of directly estimating the affine matrix. Formally, the affine registration problem is reduced to $f_\theta(F, M) = [\boldsymbol{t}, \boldsymbol{r}, \boldsymbol{s}, \boldsymbol{h}]$, where $\boldsymbol{t}, \boldsymbol{r}, \boldsymbol{s}, \boldsymbol{h} \in \mathbb{R}^3$ represent the translation, rotation, scaling and shearing parameters. Given $\mathcal{T}, \mathcal{R}, \mathcal{S}$ and $\mathcal{H}$, the resulting affine matrix $\mathcal{A}$ can be derived by a set of geometric transformation matrices via matrix multiplication as $\mathcal{A} = \mathcal{T} \cdot \mathcal{R} \cdot \mathcal{S} \cdot \mathcal{H}$, where $\mathcal{T}, \mathcal{R}, \mathcal{S}$ and $\mathcal{H}$ denote the translation, rotation, scaling and shearing transformation matrices derived by the corresponding geometric transformation parameters ($\boldsymbol{t}, \boldsymbol{r}, \boldsymbol{s}$ and $\boldsymbol{h}$), respectively. Our proposed transformation model can easily be transferred to other parametric registration settings by pruning or modifying undesired geometric transformation matrices. For instance, our C2FViT can be applied to rigid registration by removing the scaling and shearing matrices. Furthermore, our transformation model is capable of geometrical constraints, reducing the searching space of the model during optimization. In this work, the output geometric transformation parameters are constrained as follows: rotation and shearing parameters are constrained between $-\pi$ and $+\pi$, the translation parameters are constrained between -50% and +50% of the maximum spatial resolution, and the scaling parameters are constrained between 0.5 and 1.5. In this paper, we use the center of mass of the input instead of the geometric center for rotation and shearing. The center of mass c_I of the image I is defined as $c_I = \frac{\sum_{p \in \Omega} p I(p)}{\sum_{p \in \Omega} I(p)}$. If the background intensity of the image scan is non-zero, the origin of the rotation can be set to the geometric center of the image.

3.3. Unsupervised and Semi-supervised Learning

In contrast to the conventional affine registration methods, we parametrize the affine registration problem as a learning problem. Specifically, we formulate the function $f_\theta(F, M) = \mathcal{A}_f$, where f_θ and $\mathcal{A}_f$ represent the C2FViT model and the output affine transformation matrix, respectively. Mathematically, our goal is to minimize the following equation:

$$\theta^* = \arg\min_\theta \left[\mathbb{E}_{(F,M) \in D} \, \mathcal{L}\big(F, M(\phi(\mathcal{A}_f))\big) \right], \quad (2)$$

where the θ is the learning parameters in C2FViT, fixed and moving images are randomly sampled from the training dataset D and the loss function $\mathcal{L}$ measures the dissimilarity between the fixed image and the affine transformed moving image $M(\phi(\mathcal{A}_f))$. In our unsupervised learning setting, we use the negative NCC similarity measure with the similarity
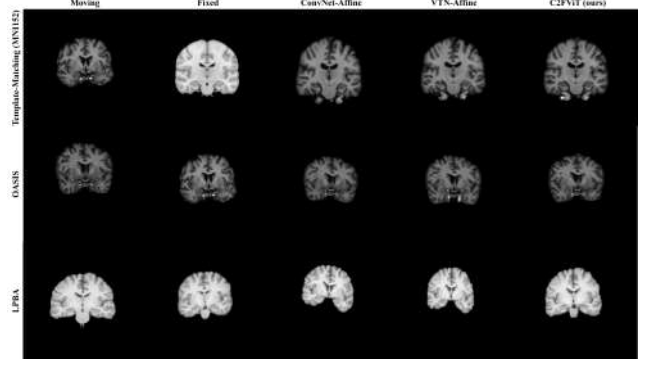

Figure 3. Example coronal MR slices from the atlases (fixed images), moving images, resulting warped images for ConvNet-Affine, VTN-Affie and our method without center of mass initialization.

pyramid [35] $\mathcal{L}_{sim}$ to quantify the distance between F and $M(\phi(\mathcal{A}_f))$ such that $\mathcal{L} = \mathcal{L}_{sim}$ and $\mathcal{L}_{sim}$ is defined as:

$$\mathcal{L}_{sim}(F, M(\phi)) = \sum_{i \in [1..L]} -\frac{1}{2^{(L-i)}} \text{NCC}_w(F_i, M_i(\phi)),$$
$$(3)$$

where L denotes the number of image pyramid levels, NCC_w represents the local normalized cross-correlation with windows size w^3, and (F_i, M_i) denotes the images in the image pyramid, *i.e.*, F_1 is the image with the lowest resolution. In addition, our method is also capable of semi-supervised learning if the anatomical segmentation maps of the fixed and moving images are available in the training dataset. Given anatomical segmentation maps of fixed image S_F and warped moving image $S_M(\phi)$, the semi-supervised C2FViT can be formulated by changing the similarity measure $\mathcal{L}$ in eq. 2 to $\mathcal{L}_{sim} + \lambda \mathcal{L}_{seg}$, where $\mathcal{L}_{seg}$ is defined as follows:

$$\mathcal{L}_{seg}(S_F, S_M(\phi)) = \frac{1}{K} \sum_{i \in [1..K]} \left(1 - \frac{2(S_F^i \cap S_M^i(\phi))}{|S_F^i| + |S_M^i(\phi)|} \right)$$
$$(4)$$

where K denotes the number of anatomical structures. For the semi-supervised C2FViT, we utilize all available anatomical segmentations in our experiments. In this paper, we employ $L = 3$ image pyramid levels and $\lambda = 0.5$.

4. Experiments

4.1. Data and Pre-processing

We evaluated our method on brain template-matching normalization and atlas-based registration using 414 T1-weighted brain MRI scans from the OASIS dataset [30] and

40 brain MRI scans from the LPBA dataset [39]. For the OASIS dataset, we resampled and padded all MRI scans to $256 \times 256 \times 256$ with the same resolution ($1mm \times 1mm \times 1mm$) followed by standard preprocessing steps, including motion correction, skull stripping and subcortical structure segmentation, for each MRI scan using FreeSurfer [14]. For the LPBA dataset, the MRI scans are skull-stripped, and the manual delineation of the subcortical structures are provided. All brain MRI scans in our experiments are in native space, except the MNI152 brain template. We split the OASIS dataset into 255, 10 and 149 volumes for training, validation, and test sets, respectively. For the LPBA dataset, we included all 40 scans as the test set.

We evaluated our method on two applications of brain registration: brain template-matching normalization to MNI152 space and atlas-based registration in native space. Brain template-matching normalization is a standard application in analyzing inter-subject images and a necessary pre-processing step in most deformable image registration methods. For the task of brain template-matching normalization, we affinely register all test scans in the OASIS dataset to an MNI152 (6th generation) brain template [13–15], which is derived from 152 structural images and averaged together after non-linear registration into the common MNI152 co-ordinate system. We train the learning-based methods with the training dataset of OASIS and the MNI152 template, which employ the MNI152 template as the fixed image and MRI scans from the training dataset as moving images. For the atlas-based registration task, we randomly select 3 and 2 scans from the test set of OASIS and LPBA datasets respectively as atlases. Then, we align the remaining MRI scans in the test set to the selected atlases within the same dataset. Note that in the atlas-based registration task, we train the learning-based methods with pairwise brain registration, which randomly samples two image scans as fixed and moving images, using only the training set of the OASIS dataset, *i.e.*, the selected atlases and the MRI scans from the LPBA dataset were not involved in the training.

Conventionally, affine registration methods often initialize the input images with center of mass (CoM) initialization by default [32], which initializes the translation parameters using the CoM of the input images. Equivalently, the CoM initialization for learning-based methods can be achieved by translating the CoM of the moving image to the CoM of the fixed image. We evaluated our method with and without the CoM initialization, and the results are listed in table 1 and table 2, respectively.

4.2. Measurement

To quantify the registration performance of an affine registration algorithm, we register each subject to an atlas or MNI152 template, propagate the subcortical structure segmentation map using the resulting affine transformation matrix, and measure the volume overlap using the Dice similarity coefficient (DSC) and 30% lowest DSC of all cases (DSC30). We also measure the 95% percentile of the Hausdorff distance (HD95) of the segmentation map to represent the reliability of the registration algorithm. In the brain template-matching normalization task, 4 subcortical structures, *i.e.*, caudate, cerebellum, putamen and thalamus, are included in the evaluation. In the atlas-based registration with the OASIS dataset, 23 subcortical structures are included, as shown in the boxplot in figure 4. For the atlas-based registration with the LPBA dataset, we utilize all manual segmentation of the brain scan, including cerebrospinal fluid (CSF), gray matter (GM) and white matter (WM), for evaluation.

4.3. Baseline Methods

We compare our method with two state-of-the-art conventional affine registration methods (ANTs [1] and Elastix [25]) and two learning-based affine registration approaches (ConvNet-Affine [11] and VTN-Affine [46]). Specifically, we use the ANTs affine registration implementation in the publicly available ANTs software package [2], and we use the Elastix affine registration algorithm in the SimpleElastix toolbox [31]. Both methods use a 3-level multi-resolution optimization strategy with adaptive gradient descent optimization and the mutual information as the similarity measure. For ConvNet-Affine and VTN-Affine, we follow their papers to implement their affine subnetworks. The initial number of feature channels for both methods is set to 16, and we follow the rules in their papers to define the growth of network depth and the hidden dimension of each convolution layer. By default, all learning-based methods are trained in an unsupervised manner with the similarity pyramid as described in eq. 3. We also extend the unsupervised learning-based methods to semi-supervised variants using the same semi-supervised object function as our method, denoted as C2FViT-semi, ConvNet-Affine-semi and VTN-Affine-semi.

4.4. Implementation

The learning-based methods, *i.e.*, C2FViT, ConvNet-Affine and VTN-Affine, are developed and trained using Pytorch. All the methods are trained or executed on a standalone workstation equipped with an Nvidia TITAN RTX GPU and an Intel Core i7-7700 CPU. The learning-based approaches are trained with half-resolution image scans by downsampling the image scans with trilinear interpolation. Then, we apply the resulting affine transformation to the full-resolution image scans for evaluation. We adopt the Adam optimizer [24] with a fixed learning rate of $1e^{-4}$ and batch size sets to 1 for all learning-based approaches.

Method	#Param	Template-Matching Normalization (MNI152)				Atlas-Based Registration (OASIS)				Atlas-Based Registration (OASIS$_{train}$ ⇒ LPBA$_{test}$)			
		DSC$_4$ ↑	DSC30$_4$ ↑	HD95$_4$ ↓	T$_{test}$ ↓	DSC$_{23}$ ↑	DSC30$_{23}$ ↑	HD95$_{23}$ ↓	T$_{test}$ ↓	DSC$_3$ ↑	DSC30$_3$ ↑	HD95$_3$ ↓	T$_{test}$ ↓
Initial	-	0.14 ± 0.12	0.02 ± 0.02	29.26 ± 11.33	-	0.18 ± 0.14	0.06 ± 0.02	15.53 ± 6.77	-	0.33 ± 0.06	0.26 ± 0.03	12.43 ± 4.65	-
ConvNet-Affine [11]	14.7 M	0.65 ± 0.08	0.56 ± 0.06	6.14 ± 1.33	0.12 ± 0.09 s	0.57 ± 0.07	0.48 ± 0.05	4.10 ± 1.01	0.09 ± 0.06 s	0.36 ± 0.07	0.28 ± 0.03	11.58 ± 4.99	0.11 ± 0.08 s
VTN-Affine [46]	14.0 M	0.67 ± 0.06	0.60 ± 0.05	5.80 ± 1.01	**2e-3** ± 4e-4 s	0.57 ± 0.08	0.48 ± 0.06	4.18 ± 1.08	**3e-3** ± 8e-4 s	0.31 ± 0.06	0.24 ± 0.03	14.99 ± 5.34	**2e-3** ± 6e-4 s
C2FViT (ours)	15.2 M	**0.71** ± 0.06	**0.64** ± 0.04	**5.17** ± 0.81	0.09 ± 0.03 s	**0.64** ± 0.06	**0.57** ± 0.05	**3.33** ± 0.77	0.08 ± 0.01 s	**0.47** ± 0.04	**0.42** ± 0.02	**6.55** ± 1.60	0.14 ± 0.06 s

Table 1. Quantitative results of template-matching normalization and atlas-based registration *without center of mass initialization*. The subscript of each metric indicates the number of anatomical structures involved. ↑: higher is better, and ↓: lower is better. Initial: initial results in native space without registration.

Method	#Param	Template-Matching Normalization (MNI152)				Atlas-Based Registration (OASIS)				Atlas-Based Registration (OASIS$_{train}$ ⇒ LPBA$_{test}$)			
		DSC$_4$ ↑	DSC30$_4$ ↑	HD95$_4$ ↓	T$_{test}$ ↓	DSC$_{23}$ ↑	DSC30$_{23}$ ↑	HD95$_{23}$ ↓	T$_{test}$ ↓	DSC$_3$ ↑	DSC30$_3$ ↑	HD95$_3$ ↓	T$_{test}$ ↓
Initial (CoM)	-	0.49 ± 0.11	0.35 ± 0.06	11.06 ± 3.48	-	0.45 ± 0.12	0.29 ± 0.06	6.97 ± 2.89	-	0.45 ± 0.04	0.41 ± 0.01	6.87 ± 1.69	-
Elastix [25]	-	0.73 ± 0.07	0.64 ± 0.06	5.01 ± 1.44	6.6 ± 0.2 s	0.63 ± 0.09	0.52 ± 0.08	3.89 ± 1.72	6.3 ± 0.2 s	**0.55** ± 0.02	**0.53** ± 0.02	4.11 ± 1.01	6.4 ± 0.2 s
ANTs [1]	-	0.74 ± 0.06	0.67 ± 0.05	4.65 ± 0.57	38.2 ± 3.2 s	0.67 ± 0.08	0.58 ± 0.08	3.27 ± 1.56	37.7 ± 2.5 s	0.54 ± 0.03	0.50 ± 0.02	4.53 ± 1.38	46.6 ± 15.3 s
ConvNet-Affine [11]	14.7 M	0.70 ± 0.06	0.63 ± 0.05	5.28 ± 0.68	0.12 ± 0.08 s	0.62 ± 0.06	0.55 ± 0.05	3.43 ± 0.91	0.10 ± 0.07 s	0.45 ± 0.04	0.41 ± 0.01	7.46 ± 1.87	0.11 ± 0.08 s
VTN-Affine [46]	14.0 M	0.71 ± 0.06	0.64 ± 0.05	5.11 ± 0.74	3e-3 ± 9e-4 s	0.66 ± 0.06	0.59 ± 0.06	3.02 ± 0.81	**2e-3** ± 7e-4 s	0.43 ± 0.04	0.39 ± 0.02	8.02 ± 2.23	**2e-3** ± 6e-4 s
C2FViT [ours]	15.2 M	0.72 ± 0.06	0.65 ± 0.05	4.99 ± 0.75	0.12 ± 0.04 s	0.66 ± 0.05	0.61 ± 0.04	2.96 ± 0.54	0.09 ± 0.02 s	0.54 ± 0.03	0.51 ± 0.04	**4.06** ± 1.12	0.12 ± 0.04 s
ConvNet-Affine-semi [11]	14.7 M	0.73 ± 0.06	0.66 ± 0.04	4.94 ± 0.76	0.12 ± 0.09 s	0.63 ± 0.06	0.56 ± 0.06	3.46 ± 0.96	0.10 ± 0.07s	0.43 ± 0.03	0.40 ± 0.02	6.90 ± 1.52	0.12 ± 0.08 s
VTN-Affine-semi [46]	14.0 M	0.75 ± 0.05	**0.70** ± 0.04	4.65 ± 0.66	**2e-3** ± 6e-4 s	0.68 ± 0.05	0.62 ± 0.04	2.94 ± 0.64	**2e-3** ± 8e-4 s	0.44 ± 0.04	0.40 ± 0.02	7.27 ± 1.96	**2e-3** ± 1e-3 s
C2FViT-semi (ours)	15.2 M	**0.76** ± 0.05	**0.70** ± 0.04	**4.60** ± 0.69	0.13 ± 0.05 s	**0.69** ± 0.04	**0.64** ± 0.04	**2.81** ± 0.55	0.08 ± 0.02 s	0.51 ± 0.03	0.47 ± 0.04	4.58 ± 1.71	0.13 ± 0.05 s

Table 2. Quantitative results on template-matching normalization, OASIS and LPBA dataset *with center of mass initialization*. The subscript of each metric indicates the number of anatomical structures involved. ↑: higher is better, and ↓: lower is better. Initial (CoM): initial results with the center of mass initialization. To our knowledge, ANTs and Elastix do not have a GPU implementation.

4.5. Results

4.5.1 Registration accuracy and Robustness

Table 1 shows the results of template-matching normalization and atlas-based registration of the learning-based methods *without spatial initialization*. Figure 3 illustrates the qualitative results of all tasks without spatial initialization. The low initial Dice scores over all subjects, suggesting that there is a large misalignment within each test case. Our proposed method is significantly better than ConvNet-Affine and VTN-Affine in terms of DSC, DSC30 and HD95 over all three tasks, suggesting our method is robust and accurate in affine registration with large initial misalignment. We visualize the distribution of Dice scores for each subcortical structure as in the boxplot in figure 4. Compared to VTN-Affine, the C2FViT model achieves consistently better performance across all structures.

Table 2 shows the results of tasks with CoM initialization. This simple but effective initialization boosts the initial Dice scores from 0.14, 0.18 and 0.33 to 0.49, 0.45 and 0.45, respectively, implying that the initialization eliminates most of the misalignment due to translation. All three learning-based methods improve significantly on affine alignment with CoM initialization. For an unsupervised manner, our method achieves comparable Dice measures to the conventional methods (ANTs and Elastix), and slightly better than ConvNet-Affine and VTN-Affine. It is worth noting that VTN-Affine gains significant improvement in registration performance of template-matching and atlas-based registration (OASIS) under CoM initialization. Nevertheless, the validity of the initial registration should

be questioned when the two images are acquired in different imaging modalities and hence, the registration performance without spatial initialization should be considered when evaluating the learning-based affine registration algorithm. With our proposed semi-supervised settings, our method C2FViT-semi achieves the best overall registration performance in the template-matching normalization and the atlas-based registration task on the OASIS dataset.

4.5.2 Generalizability Analysis

As shown in the results of the LPBA dataset in tables 1 and 2, ConvNet-Affine and VTN-Affine, using models trained on the OASIS dataset, fail spectacularly in the test set of LPBA, which obtain -5% and -2% loss in DSC with VTN-Affine, and +3% and +0% gain in DSC with ConvNet-Affine compared to initial results without registration and with spatial initialization, respectively. The results imply that their models cannot generalize well to an unseen dataset in practice regardless of spatial initialization. By contrast, our C2FViT model achieves a comparable registration performance to the conventional affine registration approaches ANTs and Elastix in the task with the LPBA dataset, reaching an average Dice score of 0.54 and HD95 of 4.06 in the task with the LPBA dataset, as shown in table 2. While the semi-supervised settings improve the dataset-specific performance of learning-based models in template-matching normalization and atlas-based registration with the OASIS dataset, the semi-supervised models are inferior to their unsupervised models in the LPBA dataset, indicating anatomical knowledge injected to the model with semi-supervision may not generalize well to unseen data beyond the training dataset.
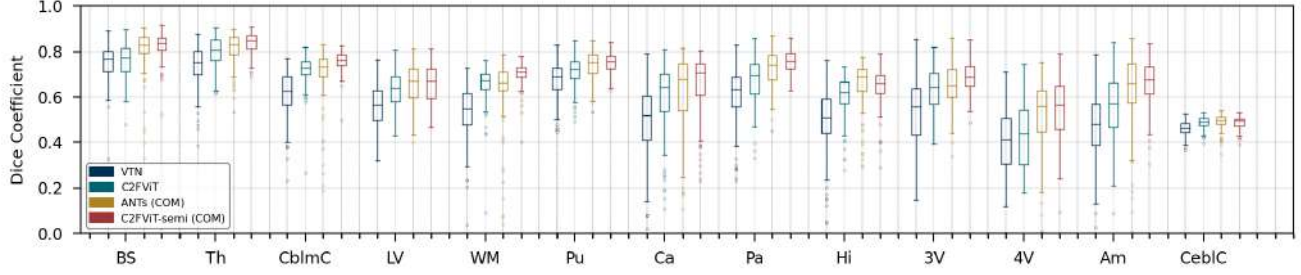
Figure 4. Boxplots illustrating Dice scores of each anatomical structure for C2FViT, VTN and ANTs in the atlas-based registration with the OASIS dataset. The left and right hemispheres of the brain are combined into one structure for visualization. The brain stem (BS), thalamus (Th), cerebellum cortex (CblmC), lateral ventricle (LV), cerebellum white matter (WM), putamen (Pu), caudate (Ca), pallidum (Pa), hippocampus (Hi), 3rd ventricle (3V), 4th ventricle (4V), amygdala (Am), and cerebral cortex (CeblC) are included. Methods with (CoM) postfix are trained and tested on MRI scans with the center of mass initialization.

Methods	DSC_{23}	$HD95_{23}$	T_{test}	#Param
Vanilla C2FViT-s1	0.61	3.53	0.05 ± 0.04 s	5.0 M
Vanilla C2FViT-s2	0.62	3.57	0.06 ± 0.05 s	10.0 M
Vanilla C2FViT-s3	0.62	3.46	0.07 ± 0.02 s	15.2 M
+Progressive Spatial Transformation	0.64 (+0.02)	3.33 (-0.13)	0.08 ± 0.02 s	15.2 M
+Center of Mass Initialization	0.66 (+0.02)	2.96 (-0.37)	0.09 ± 0.02 s	15.2 M
+Semi-supervision	0.69 (+0.03)	2.81 (-0.15)	0.08 ± 0.02 s	15.2 M

Table 3. Influence of the number of stages, progressive spatial transformation, center of mass initialization and the semi-supervised learning to the C2FViT model. The C2FViT with postfix -s$\{n\}$ represents the C2FViT model with an n-stage.

4.5.3 Runtime Analysis

The average runtimes (denoted as T_{test}) of all methods in the inference phase are reported in tables 1 and 2. We report the average registration time for each task. C2FViT, ConvNet-Affine and VTN-Affine are faster than the ANTs and Elastix by order of magnitude, thanks to the GPU acceleration and the effective learning formulation. Moreover, ANTs runtimes vary widely, as its convergence depends on the degree of initial misalignment of the task. On the other hand, Elastix runtimes are stable at around 6.6 seconds per alignment task because of the early stopping strategy used during the affine alignment.

	$DSC_{23} \uparrow$	$DSC30_{23} \uparrow$	$HD95_{23} \downarrow$	$T_{test} \downarrow$
C2FViT-direct	0.63 ± 0.06	0.55 ± 0.04	3.43 ± 0.73	$0.02 \pm 4e\text{-}3$ s
C2FViT-decouple	0.64 ± 0.06	0.57 ± 0.05	3.33 ± 0.77	0.08 ± 0.01 s

Table 4. Influence of the proposed decoupled affine transmation model compared to the direct affine matrix estimation model.

4.5.4 Ablation study

Table 3 shows the ablation study results of C2FViT in the OASIS atlas-based registration task. The results suggest that the proposed progressive spatial transformation, CoM initialization and semi-supervised learning consistently improve the registration performance of C2FViT without adding extra learning parameters or significant computational burden to the model. Table 4 presents the results of C2FViT using two different transformation models in the OASIS atlas-based registration task. The proposed decoupled affine transformation model is slightly better than directly learning the affine matrix, in terms of registration performance, at the cost of registration runtime. Moreover, the decoupled affine transformation model can be easily adapted to other parametric registration methods by pruning or modifying the geometrical transformation matrices.

5. Conclusion

We have proposed a Coarse-to-Fine Vision Transformer dedicated to 3D affine medical image registration. Unlike prior works using CNN-based affine registration methods, our method leverages the global connectivity of the self-attention operator and moderates the locality of the convolutional feed-forward layer to encode the global orientations, spatial positions and long-term dependencies of the image pair to a set of geometric transformation parameters. Comprehensive experiments demonstrate that our method not only achieves superior registration performance over the existing CNN-based methods under data with large initial misalignment and is robust to an unseen dataset, but also our method with semi-supervision outperforms conventional methods in terms of dataset-specific and preserves the runtime advantage of learning-based methods. Nevertheless, there is still a gap between unsupervised learning-based approaches and conventional approaches. We believe that expanding the training dataset and introducing task-specific data augmentation techniques would likely lead to performance improvement.

References

[1] Brian B Avants, Nick Tustison, and Gang Song. Advanced normalization tools (ANTS). *Insight j*, 2(365):1–35, 2009. 2, 6, 7

[2] Brian B Avants, Nicholas J Tustison, Gang Song, and Others. A reproducible evaluation of ANTs similarity metric performance in brain image registration. *Neuroimage*, 54(3):2033–2044, 2011. 6

[3] Guha Balakrishnan, Amy Zhao, Mert R Sabuncu, John Guttag, and Adrian V Dalca. An unsupervised learning model for deformable medical image registration. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9252–9260, 2018. 1

[4] Junyu Chen, Yufan He, Eric C Frey, Ye Li, and Yong Du. Vit-v-net: Vision transformer for unsupervised volumetric medical image registration. *Medical Imaging with Deep Learning*, 2021. 3

[5] Xu Chen, Yanda Meng, Yitian Zhao, Rachel Williams, Srinivasa R Vallabhaneni, and Yalin Zheng. Learning unsupervised parameter-specific affine transformation for medical images registration. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 24–34. Springer, 2021. 2

[6] Zhengsu Chen, Lingxi Xie, Jianwei Niu, Xuefeng Liu, Longhui Wei, and Qi Tian. Visformer: The vision-friendly transformer. *arXiv preprint arXiv:2104.12533*, 2021. 3

[7] Xiangxiang Chu, Zhi Tian, Yuqing Wang, Bo Zhang, Haibing Ren, Xiaolin Wei, Huaxia Xia, and Chunhua Shen. Twins: Revisiting the design of spatial attention in vision transformers. *arXiv preprint arXiv:2104.13840*, 1(2):3, 2021. 3

[8] Zihang Dai, Hanxiao Liu, Quoc V Le, and Mingxing Tan. Coatnet: Marrying convolution and attention for all data sizes. *arXiv preprint arXiv:2106.04803*, 2021. 3

[9] Adrian V Dalca, Guha Balakrishnan, John Guttag, and Mert R Sabuncu. Unsupervised learning for fast probabilistic diffeomorphic registration. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 729–738. Springer, 2018. 1

[10] Stéphane d'Ascoli, Hugo Touvron, Matthew Leavitt, Ari Morcos, Giulio Biroli, and Levent Sagun. Convit: Improving vision transformers with soft convolutional inductive biases. *arXiv preprint arXiv:2103.10697*, 2021. 2, 3

[11] Bob D de Vos, Floris F Berendsen, Max A Viergever, Hessam Sokooti, Marius Staring, and Ivana Išgum. A deep learning framework for unsupervised affine and deformable image registration. *Medical image analysis*, 52:128–143, 2019. 1, 2, 6, 7

[12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2, 3, 4

[13] Alan C Evans, Andrew L Janke, D Louis Collins, and Sylvain Baillet. Brain templates and atlases. *Neuroimage*, 62(2):911–922, 2012. 2, 6

[14] Bruce Fischl. FreeSurfer. *Neuroimage*, 62(2):774–781, 2012. 2, 6

[15] GŘnther Grabner, Andrew L Janke, Marc M Budge, David Smith, Jens Pruessner, and D Louis Collins. Symmetric atlasing and model based segmentation: an application to the hippocampus in older adults. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 58–66. Springer, 2006. 2, 6

[16] Jianyuan Guo, Kai Han, Han Wu, Chang Xu, Yehui Tang, Chunjing Xu, and Yunhe Wang. Cmt: Convolutional neural networks meet vision transformers. *arXiv preprint arXiv:2107.06263*, 2021. 3

[17] Mattias P Heinrich. Closing the gap between deep and conventional image registration using probabilistic dense displacement networks. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 50–58. Springer, 2019. 1

[18] Mattias P Heinrich, Oskar Maier, and Heinz Handels. Multimodal Multi-Atlas Segmentation using Discrete Optimisation and Self-Similarities. *VISCERAL Challenge@ ISBI*, 1390:27, 2015. 2

[19] Alessa Hering, Stephanie Häger, Jan Moltz, Nikolas Lessmann, Stefan Heldmann, and Bram van Ginneken. Cnn-based lung ct registration with multiple anatomical constraints. *Medical Image Analysis*, page 102139, 2021. 1

[20] Andrew Hoopes, Malte Hoffmann, Bruce Fischl, John Guttag, and Adrian V Dalca. Hypermorph: Amortized hyperparameter learning for image registration. In *International Conference on Information Processing in Medical Imaging*, 2021. 1

[21] Yipeng Hu, Marc Modat, Eli Gibson, Nooshin Ghavami, Ester Bonmati, Caroline M Moore, Mark Emberton, J Alison Noble, Dean C Barratt, and Tom Vercauteren. Label-driven weakly-supervised learning for multimodal deformable image registration. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 1070–1074. IEEE, 2018. 1, 2, 4

[22] Weijian Huang, Hao Yang, Xinfeng Liu, Cheng Li, Ian Zhang, Rongpin Wang, Hairong Zheng, and Shanshan Wang. A coarse-to-fine deformable transformation framework for unsupervised multi-contrast mr image registration with dual consistency constraint. *IEEE Transactions on Medical Imaging*, 2021. 1, 2

[23] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and Others. Spatial transformer networks. In *Advances in neural information processing systems*, pages 2017–2025, 2015. 4

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

[25] Stefan Klein, Marius Staring, Keelin Murphy, Max A Viergever, and Josien PW Pluim. Elastix: a toolbox for intensity-based medical image registration. *IEEE transactions on medical imaging*, 29(1):196–205, 2009. 2, 6, 7

[26] Yawei Li, Kai Zhang, Jiezhang Cao, Radu Timofte, and Luc Van Gool. Localvit: Bringing locality to vision transformers. *arXiv preprint arXiv:2104.05707*, 2021. 3, 4

[27] Min Lin, Qiang Chen, and Shuicheng Yan. Network in network. *arXiv preprint arXiv:1312.4400*, 2013. 2

[28] Rosanne Liu, Joel Lehman, Piero Molino, Felipe Petroski Such, Eric Frank, Alex Sergeev, and Jason Yosinski. An intriguing failing of convolutional neural networks and the coordconv solution. *arXiv preprint arXiv:1807.03247*, 2018. 2

[29] JB Antoine Maintz and Max A Viergever. A survey of medical image registration. *Medical image analysis*, 2(1):1–36, 1998. 1

[30] Daniel S Marcus, Tracy H Wang, Jamie Parker, John G Csernansky, John C Morris, and Randy 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(9):1498–1507, 2007. 5

[31] Kasper Marstal, Floris Berendsen, Marius Staring, and Stefan Klein. Simpleelastix: A user-friendly, multi-lingual library for medical image registration. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 134–142, 2016. 6

[32] Matthew Michael McCormick, Xiaoxiao Liu, Luis Ibanez, Julien Jomier, and Charles Marion. ITK: enabling reproducible research and open science. *Frontiers in neuroinformatics*, 8:13, 2014. 6

[33] Shun Miao, Z Jane Wang, and Rui Liao. A cnn regression approach for real-time 2d/3d registration. *IEEE transactions on medical imaging*, 35(5):1352–1363, 2016. 2

[34] Tony C W Mok and Albert Chung. Fast Symmetric Diffeomorphic Image Registration with Convolutional Neural Networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4644–4653, 2020. 1

[35] Tony C W Mok and Albert C S Chung. Large Deformation Diffeomorphic Image Registration with Laplacian Pyramid Networks. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 211–221. Springer, 2020. 1, 5

[36] Tony C W Mok and Albert C S Chung. Conditional Deformable Image Registration with Convolutional Neural Network. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2021. 1

[37] Josien PW Pluim, JB Antoine Maintz, and Max A Viergever. Mutual-information-based registration of medical images: a survey. *IEEE transactions on medical imaging*, 22(8):986–1004, 2003. 1

[38] Wei Shao, Indrani Bhattacharya, Simon JC Soerensen, Christian A Kunder, Jeffrey B Wang, Richard E Fan, Pejman Ghanouni, James D Brooks, Geoffrey A Sonn, and Mirabela Rusu. Weakly supervised registration of prostate mri and histopathology images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 98–107. Springer, 2021. 2, 4

[39] David W Shattuck, Mubeena Mirza, Vitria Adisetiyo, Cornelius Hojatkashani, Georges Salamon, Katherine L Narr, Russell A Poldrack, Robert M Bilder, and Arthur W Toga. Construction of a 3D probabilistic atlas of human cortical structures. *Neuroimage*, 39(3):1064–1080, 2008. 6

[40] Zhengyang Shen, Xu Han, Zhenlin Xu, and Marc Niethammer. Networks for joint affine and non-parametric image registration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4224–4233, 2019. 1, 2

[41] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 2, 3, 4

[42] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pvtv2: Improved baselines with pyramid vision transformer. *arXiv preprint arXiv:2106.13797*, 2021. 3, 4

[43] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *IEEE ICCV*, 2021. 2, 3

[44] Haiping Wu, Bin Xiao, Noel Codella, Mengchen Liu, Xiyang Dai, Lu Yuan, and Lei Zhang. Cvt: Introducing convolutions to vision transformers. *arXiv preprint arXiv:2103.15808*, 2021. 2, 3

[45] Yungeng Zhang, Yuru Pei, and Hongbin Zha. Learning dual transformer network for diffeomorphic registration. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 129–138. Springer, 2021. 3

[46] Shengyu Zhao, Tingfung Lau, Ji Luo, I Eric, Chao Chang, and Yan Xu. Unsupervised 3d end-to-end medical image registration with volume tweening network. *IEEE journal of biomedical and health informatics*, 24(5):1394–1404, 2019. 1, 2, 4, 6, 7

[47] Wu Zhou, Lijuan Zhang, Yaoqin Xie, and Changhong Liang. A novel technique for prealignment in multimodality medical image registration. *BioMed research international*, 2014, 2014. 1, 2

A. Unsupervised and Semi-Supervised Learning

Figure 5 depicts the proposed unsupervised and semi-supervised training scheme of the Coarse-to-Fine Vision Transformer (C2FViT). The segmentation maps are only required in the training phrase under the semi-supervised training scheme.
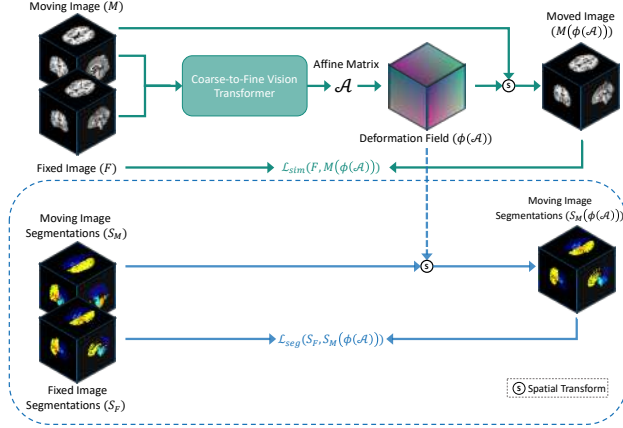

Figure 5. Schematic representation of the unsupervised and semi-supervised learning scheme in the Coarse-to-Fine Vision Transformer. The unsupervised and semi-supervised learning schemes are highlighted in green and blue colours, respectively.

B. Affine Transformations

The corresponding translation $\mathcal{T}$, rotation $\mathcal{R}$, scaling $\mathcal{S}$ and shearing $\mathcal{H}$ transformations derived by the geometric transformation parameters $t_x, t_y, t_z \in \boldsymbol{t}$, $r_x, r_y, r_z \in \boldsymbol{r}$, $s_x, s_y, s_z \in \boldsymbol{s}$ and $h_x, h_y, h_z \in \boldsymbol{h}$ are defined as follows:

$$\mathcal{T} = \begin{pmatrix} 1 & 0 & 0 & t_x \\ 0 & 1 & 0 & t_y \\ 0 & 0 & 1 & t_z \\ 0 & 0 & 0 & 1 \end{pmatrix}, \mathcal{R}_x = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & \cos(r_x) & \sin(r_x) & 0 \\ 0 & -\sin(r_x) & \cos(r_x) & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

$$\mathcal{S} = \begin{pmatrix} s_x & 0 & 0 & 0 \\ 0 & s_y & 0 & 0 \\ 0 & 0 & s_z & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \mathcal{R}_y = \begin{pmatrix} \cos(r_y) & 0 & -\sin(r_y) & 0 \\ 0 & 1 & 0 & 0 \\ \sin(r_y) & 0 & \cos(r_y) & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

$$\mathcal{H} = \begin{pmatrix} 1 & h_{xy} & h_{xz} & 0 \\ 0 & 1 & h_{yz} & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \mathcal{R}_z = \begin{pmatrix} \cos(r_z) & -\sin(r_z) & 0 & 0 \\ \sin(r_z) & \cos(r_z) & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

where rotation matrix $\mathcal{R}$ equals to $\mathcal{R}_x \mathcal{R}_y \mathcal{R}_z$.

C. Additional Implementation Details

Table 5 summarizes the configurations of C2FViT at each stage. Specifically, the input resolution, stride in the convolutional patch embedding, number of transformer encoders, embedding size of each patch embedding, embedding size of the convolutional feed-forward layer and num-ber of heads for the multi-head self-attention module are listed in the table.

Stage	Input size	Stride	# Encoders	Hidden size	MLP size	Heads
Stage 1	32^3	2^3	4	256	512	2
Stage 2	64^3	4^3	4	256	512	2
Stage 3	128^3	8^3	4	256	512	2

Table 5. Model configurations of Coarse-to-Fine Vision Transformer at each stage.

D. Additional Qualitative Results

Figure 6 shows example MR slices obtained from the MNI152 template, OASIS and LPBA datasets. As shown in the figure, there are significant spatial and structural differences across scans as all scans are in native space, except for the MNI152 template. The comprehensive qualitative results of template-matching normalization and atlas-based registration tasks with the OASIS and LPBA dataset of the learning-based methods *without spatial initialization* are shown in figure 7.
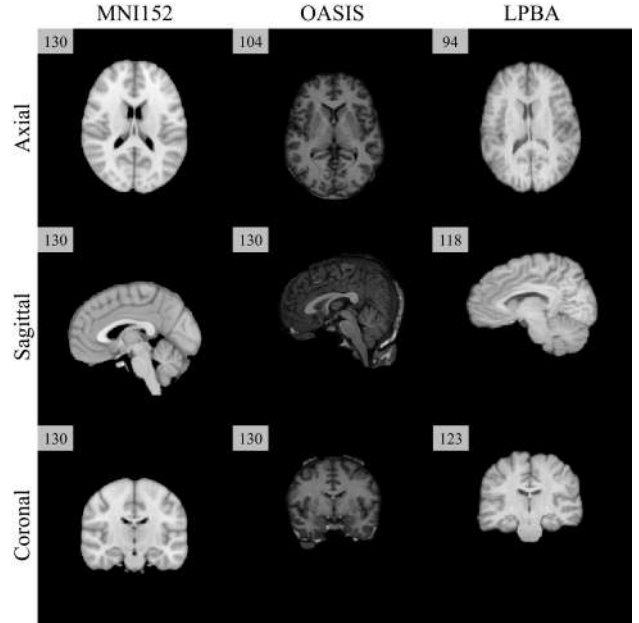

Figure 6. Example axial, sagittal and coronal slices obtained from the MNI152 template, OASIS and LPBA brain MRI datasets. The corresponding slice number of each slice is highlighted at the top-left corner.

E. Details of ANTs and Elastix

The command and parameters we used for ANTs:

```
-d 3 -v 1 -t Affine[0.1]
-m MI[<Fixed>,<Moving>,1,32,Regular,0.1]
-c 200x200x200 -f 4x2x1 -s 2x1x0
-o <OutFileSpec>
```

The command and parameters we used for Elastix:

```
ef = sitk.ElastixImageFilter()
ef.SetFixedImage(sitk.ReadImage(<Fixed>))
ef.SetMovingImage(sitk.ReadImage(<Moving>))
pmap = sitk.GetDefaultParameterMap("affine")
ef.SetParameterMap(pmap)
ef.Execute()
```
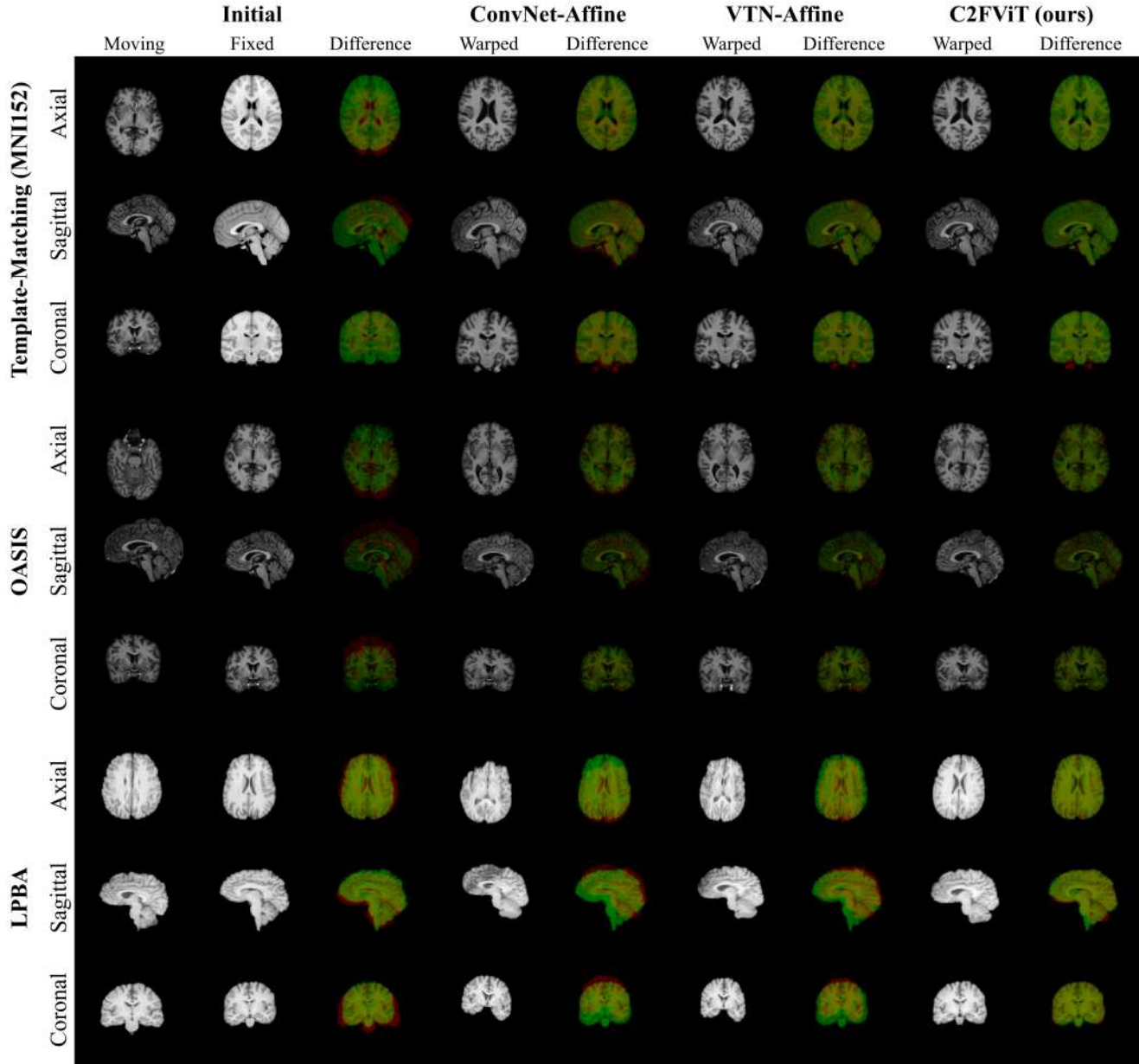
Figure 7. Example axial, sagittal and coronal MR slices obtained from the moving images, atlases (fixed images), resulting warped images for ConvNet-Affine, VTN-Affie and our method without center of mass initialization. For better visualization, we depict a difference map for each method, in which the colour maps of fixed and warped moving images are set to black-green and black-red, respectively, and overlay the resulting warped moving image to fixed image.