DOI:10.3969/j.issn.1005-202X.2025.03.004

医学影像物理

基于自相似性上下文和混合注意力的无监督可变形医学图像配准

李碧草[1],王岩[1],王贝[2],邵珠宏[3],郭旭伟[4],衣本泽[1]
1.中原工学院信息与通信工程学院,河南 郑州 450007;2.中原工学院校医院,河南 郑州 451191;3.首都师范大学信息工程学院,北京 100048;4.河南科技大学第一附属医院儿科,河南 洛阳 471000

【摘要】为了充分利用Transformer进行精确的配准,采用自相似性上下文作为特征提取器提取体素邻域上下文的语义信息。它使用具有扩散正则化的对称多尺度离散优化来寻找平滑的变换,可以快速地计算描述符之间的逐点距离。此外,提出一种基于混合注意力的Transformer网络(STWA),结合通道、空间注意力以及基于(移动)窗口的自注意力方案,充分利用3种注意力机制的互补优势,既能利用全局统计信息,又具有强大的局部拟合能力。在LPBA40、IXI和OASIS 3个3D大脑MRI数据集上进行全面的实验,结果表明,与常用的配准方法SyN、VoxelMorph、CycleMorph、ViT-V-Net和TransMorph相比,本文方法在评估指标上实现优越的性能,证明模型在可变形医学图像配准中的有效性。

【关键词】可变形医学图像配准;自相似性上下文;混合注意力;无监督深度学习
【中图分类号】R318;TP391 【文献标志码】A 【文章编号】1005-202X(2025)03-0305-08

Unsupervised deformable medical image registration based on self-similarity context and mixed attention

LI Bicao[1], WANG Yan[1], WANG Bei[2], SHAO Zhuhong[3], GUO Xuwei[4], YI Benze[1]
1. School of Electronic and Information Engineering, Zhongyuan University of Technology, Zhengzhou 450007, China; 2. Infirmary, Zhongyuan University of Technology, Zhengzhou 451191, China; 3. Information Engineering College, Capital Normal University, Beijing 100048, China; 4. Department of Pediatrics, the First Affiliated Hospital of He'nan University of Science and Technology, Luoyang 471000, China

Abstract: To fully exploit Transformer for accurate registration, self-similarity context is used as a feature extractor to extract the semantic information of the voxel neighborhood context, using symmetric multi-scale discrete optimization with diffusion regularization to find smooth transformations for quickly calculating the point-by-point distance between descriptors. In addition, a spatial-channel Transformer based on window attention network is proposed, which combines channel, spatial attention and self-attention scheme based on (moving) window, and makes full use of the complementary advantages of these 3 attention mechanisms, enabling the network to utilize global statistical information and have strong local fitting ability. The results of comprehensive experiments on 3D brain MRI datasets of LPBA40, IXI and OASIS shows that the proposed method is superior to the commonly used registration methods (SyN, VoxelMorph, CycleMorph, ViT-V-Net and TransMorph) on several evaluation indicators, proving its effectiveness in deformable medical image registration.

Keywords: deformable medical image registration; self-similarity context; mixed attention; unsupervised deep learning

前 言

可变形图像配准一直是医学影像领域的一个重要焦点,对术前规划、术中信息融合、疾病诊断和手术导航至关重要[1-3]。它的目的是对取自不同时间、不同传感器或不同视角的同一场景的两幅或多幅图像进行匹配,通过寻找一种(或一系列)空间变换,使两幅图像的对应点达到空间位置和解剖结构上的完全一致,这些图像对被称为运动图像和固定图像。设$I_m, I_f \in R^{H\times W\times L}$为运动图像和固定图像[4](H、W、L表示图像大小),在基于深度学习的配准框架中,通常需要使用空间变换网络[5]将估计的采样网格$G \in R^{H\times W\times L\times 3}$应用到运动图像,其中G是通过将规则

【收稿日期】2024-08-22
【基金项目】国家自然科学基金(61901537);河南省高校科技创新人才支持计划(23HASTIT030);中原工学院学科青年硕导培育计划(SD202207);中原工学院研究生科研创新计划(YKY2024ZK26)
【作者简介】李碧草,博士,副教授,硕士生导师,研究方向:人工智能与医学图像处理,E-mail: lbc@zut.edu.cn

网格和变形场叠加而获得的。对于采样网格中的任何位置$p \in R^3$，$G(p)$表示空间对应关系，这意味着固定图像中位置p处的体素对应于运动图像中位置$G(p)$处的体素。图像配准可以理解为在运动图像和固定图像之间找到相应的体素，并将其转换为体素之间的相对位置关系，这与Transformer计算方法非常相似[6]。

传统基于特征匹配的配准方法大多通过直观明确的特征匹配来建立图像对之间的空间对应关系[7-8]。然而，用于匹配的特征通常是手动设计的。此外，这些方法在实践中很耗时，因为它们采用基于迭代的优化方法求解一个能量函数来得到两幅图像的最优变换。最近，基于深度学习的模型，特别是卷积神经网络（CNN），已经成为许多计算机视觉任务的主流解决方案，例如图像分类、分割和检测，其强大的自动特征提取能力已经在这些任务中得到验证。后来，一些研究人员成功地将基于CNN的模型引入图像配准，解决传统方法耗时的问题[9]。基于CNN的方法在训练阶段用单个全局函数优化取代传统方法中昂贵的每幅图像优化。与传统方法相比，基于CNN的方法可以显著提高配准性能，同时在训练后运算速度更快。CNN从训练图像中学习图像配准的通用表示，从而能够在训练后快速对齐未见过的图像对。最初，为了训练神经网络，需要对金标准变形场（通常使用传统的配准方法生成）进行监督，以训练神经网络[10-12]。最近，研究重点已转向开发不依赖于金标准变形场的无监督方法[13-14]。尽管CNN在特征提取中的作用是显而易见的，但在这个过程中没有看到任何明显的特征匹配迹象。

由于卷积运算的固有局部性（即有限的有效感受野），CNN架构通常在建模图像中存在的显式长程空间关系（即相距遥远的两个体素之间的关系）方面存在局限性[15]。Transformer最初是为了解决自然语言处理领域中的问题而被提出和设计的，其关键思想是自注意力机制，模型通过它可以在处理输入序列时为每个位置分配不同的注意力权重，从而更好地捕捉长距离依赖关系[16]。这种注意力机制的并行性使得Transformer能够更高效地处理序列数据，并在短时间内取得显著的性能提升。最近，Transformer已成功应用于计算机视觉领域，在医学图像配准中也产生深远的影响[17]。Transformer可以更好地理解运动图像和固定图像之间的空间对应关系。配准是通过比较运动到固定图像的不同部分来直观地建立这种对应关系的过程。因此，将Transformer应用于图像配准的研究成为最近的焦点。Transformer具有较强的局部信息建模能力，但其利用信息的范围需要进一步扩展。在用于图像配准的U形模型中，图像被分割成1/4分辨率级的块，并被输入到Transformer中。因此，Transformer的输出表示失去了对密集变形场预测至关重要的细粒度空间信息。而且，现有的研究使用基于线性投影和特征处理的Transformer模块，这些模块缺乏空间和局部上下文来细化器官边界[18]。

为了解决上述限制并进一步开发Transformer在配准中的潜力，本文提出一种基于自相似性上下文（Self-Simialrity Context, SSC）的可变形配准网络，充分利用Transformer结构的固有变形估计能力，以实现更好的配准性能。在3个公共大脑磁共振成像（MRI）数据集上的实验表明，本文方法胜过几个顶尖的配准网络和Transformers。这项工作的主要贡献如下：（1）提出使用SSC[19]去提取体素邻域上下文细粒度的语义信息，可以非常快速地计算描述符之间的逐点距离；（2）设计一种新的基于混合注意力的Transformer网络（Spatial-channel Transformer based on Window Attention, STWA），它结合窗口注意力、通道注意力（CA）以及空间注意力，从而利用各自全局统计和强大的局部拟合能力的互补优势激活更多的像素进行更好的配准；（3）在3个公共大脑MRI数据集上进行的广泛实验表明，提出的方法优于当前最常用的配准网络和Transformers，证明本文方法在非刚性变形估计问题上的潜力。

1 方 法

1.1 网络构造

本文提出的可变形配准网络框架的主要组件包括SSC特征提取器、基于混合注意力机制的编码器、基于上采样和卷积的解码器、4个跳跃连接操作，以及用于运动图像变形的空间变换网络。整个框架的概述如图1所示，其中SSC特征提取器提取体素邻域上下文的语义信息，编码器和解码器之间的跳跃连接用于特征融合。首先，将输入的运动图像和固定图像在通道的维度上进行拼接，然后经过SSC提取每个体素邻域上下文的语义信息，随后SSC特征提取器提取到的含有语义信息的特征图被网络的编码器分割成不重叠的3D图像块，每个图像块的大小为$2\times P\times P\times P$，其中$P$通常设置为4[20-21]。将第$i$个图像块表示为$x_p^i$，$i \in \{1,\cdots,N\}$，$N = H/P \times W/P \times L/P$是patch的总数。每个图像块被展平并被视为“标记”，然后使用线性投影层将每个标记投影到任意维度的特征表示（表示为C）：

$$z_0 = \left\{x_p^1 E; x_p^2 E; \cdots; x_p^N E\right\} \tag{1}$$

其中，$E \in R^{2P^3\times C}$表示线性投影，输出z_0的维数为$N\times C$。

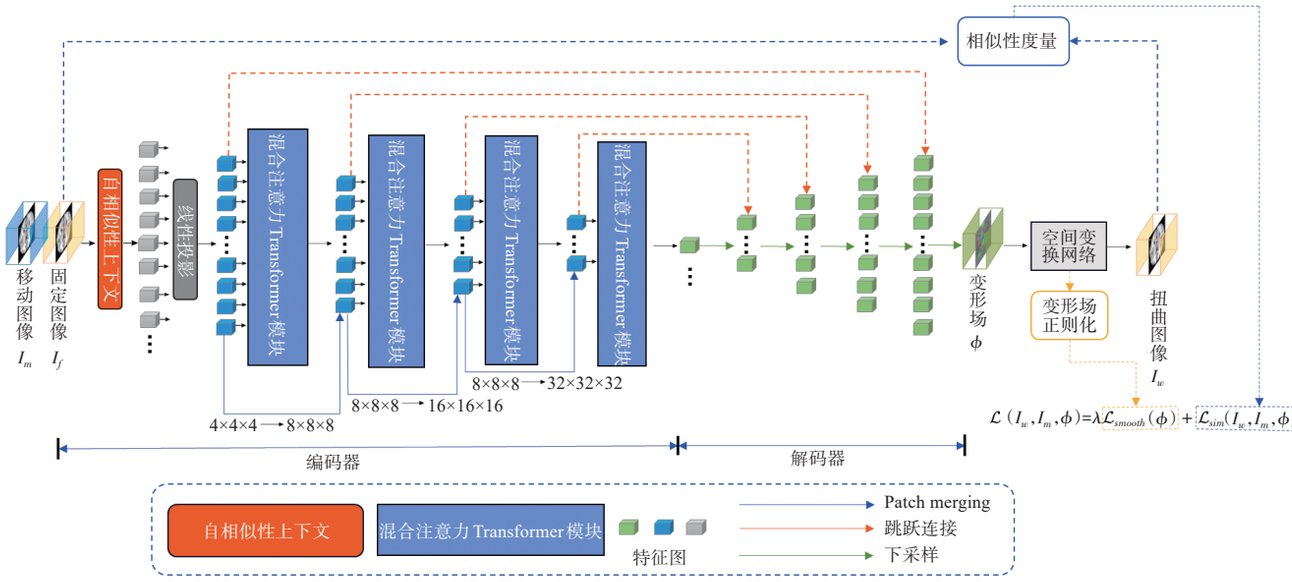

图1 网络整体构架

Figure 1 Overall network architecture

线性投影层之后，在z_0上应用patch merging和混合注意力Transformer模块的几个连续阶段。混合注意力Transformer模块输出与输入相同数量的token，而patch merging层连接每组2×2×2=8个相邻token的特征，因此它们将token数量减少2×2×2=8的倍数。然后，在8C维的级联特征上应用线性层以产生每个2C维的特征。经过4个阶段的混合注意力Transformer模块和3个阶段的Transformer阶段之间的patch merging操作（即图1中的蓝色线），编码器最后一阶段的输出尺寸为$H/32 \times W/32 \times L/32 \times 8C$。解码器由核大小为3×3的连续上采样和卷积层组成。解码阶段中的每个上采样特征图通过跳跃连接与来自编码路径的相应特征图连接，然后是两个连续的卷积层，如图1所示。由于图像块操作的性质，Transformer编码器只能提供高达$H/P \times W/P \times L/P$分辨率的特征图。因此，Transformer可能无法提供高分辨率的特征图和在较低层聚合局部信息[22]。为了解决这个缺点，使用两个卷积层，使用原始和下采样的图像对作为输入来捕获局部信息并生成高分辨率特征图。这些层的输出与解码器中的特征图连接，以产生变形场。输出变形场ϕ是通过应用16个3×3卷积生成的。除了最后一个卷积层，每个卷积层后面都有一个泄露修正线性单元激活函数（Leaky ReLU）。最后，使用空间变换函数对运动图像I_m进行非线性扭曲，扭曲所需的变形场ϕ（或位移场u）由网络提供，最终生成扭曲图像$I_m \circ \phi$。

1.2 SSC特征提取器

SSC是模态无关邻域描述符（Modality Independent Neighborhood Descriptor, MIND）的改进，它重新定义邻域布局以提高匹配的鲁棒性[23]。它还带有有效的量化方案，允许使用汉明权重（Hamming Weight）计算成对距离。SSC是通过基于图像块的自相似性方式进行估计，类似于MIND，但其目的不是提取局部形状或几何形状的表示，而是找到感兴趣体素周围的上下文。自相似性可以通过一个图像内图像块之间的距离函数（可以在同一扫描中使用平方差之和SSD）、局部或全局噪声估计σ^2以及计算自相似性的特定邻域N来描述。对于以x为中心的图像块，自相似性描述符由式（2）给出：

$$S(I,x,y) = \exp\left(-\frac{\mathrm{SSD}(x,y)}{\sigma^2}\right);\ x,y \in N \tag{2}$$

其中，y定义N内的图像块的中心位置。图2显示了自相似性上下文的结构概念。用橙色和蓝色立方体表示中心和六邻域的图像块，用黑色线连接需要计算距离的成对的边。SSC可以使用所有图像块到图像块的距离更好地描述几何和结构背景，这些距离都不依赖于中心图像块。在MIND中，邻域布局被定义为始终包括以x为中心的图像块，用于成对距离计算。中心图像块内的图像伪影或噪声总是对自相似性描述符具有直接的不利影响。因此，SSC使用6个邻域内图像块的所有成对距离来计算该位置的描述符（它们之间的欧几里得距离为$\sqrt{2}$）。式（2）中的噪声估计被定义为所有图像块距离的平均值。描述符被归一化，使得max(S)=1。

提取两幅图像的描述符，产生每个体素的向量，可以将两个图像I_f和I_m中的位置x_f和x_m的相似性度量定义为它们对应的描述符之间的绝对差之和（SAD）。因此，两个描述符之间的距离D为：

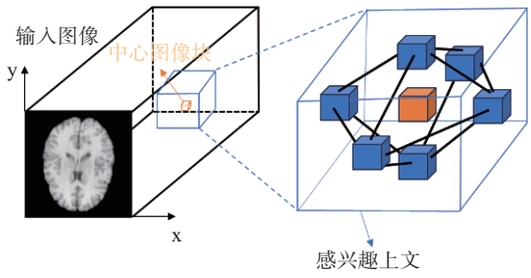

图2 3D脑部图像中中心图像块的自相似性上下文表示(橙色表示中心图像块,蓝色表示其周围上下左右前后六邻域图像块)

Figure 2 Self-similarity context representation of the central image block in the 3D brain image, with orange representing the central image block and blue representing the 6 neighborhood image blocks around it

$$D(x_f,x_m)=\frac{1}{|N|}\sum\left|S(I,x_f,y)-S(J,x_m,y)\right| \tag{3}$$

式(3)需要$|N|$次计算来评估一个体素上的相似性。这里使用的是离散化框架,每个体素使用多次成本函数评估。为了加快计算速度,将描述符量化为具有64位(bits)的单个整数值,而没有显著的精度损失,并且通过计算两个描述符之间的汉明距离可以实现在每个体素上只执行一次操作就可以获得式(2)的准确相似性评估。一个使用SSC的描述符包含12个元素,每个元素使用5位,这相当于有6个不同的可能值(注意不能使用2^5的量化,因为汉明权重只计算不同的比特数),它也可以用于其他基于多特征的配准技术。

1.3 混合注意力Transformer

为了减少模型对冗余信息的关注,提高模型对更大范围的上下文感知,本文提出混合注意力Transformer,由通道空间注意力和窗口注意力两部分组成。其中,通道空间注意力又包括通道注意力和空间注意力。通道空间注意力使模型重点关注输入特征中的重要通道以及输入数据的特定空间区域。此外,许多工作表明,卷积可以帮助Transformer获得更好的视觉表示或实现更容易的优化[24-26]。因此,本文在标准Transformer模块中加入基于空间通道注意力的卷积块,以增强网络的表示能力。如图3所示,通道空间注意力块(CSA)与基于窗口的多头自注意(W-MSA)模块并行地插入到第一层层归一化(LN)之后的标准Swin-Transformer块中。在类似于Swin-Transformer的连续混合注意力Transformer模块中,每隔一段时间就会采用基于移位窗口的自注意(SW-MSA)。为了避免CSA和MSA在优化和视觉表示方面可能发生的冲突,将一个小常数乘以CSA的输出。对于给定的输入特征x:

$$X_N=\mathrm{LN}(x) \tag{4}$$

$$X_M=(S)W-\mathrm{MSA}(X_N)+a\mathrm{CSA}(x)+x \tag{5}$$

$$Y=\mathrm{MLP}[\mathrm{LN}(X_M)]+X_M \tag{6}$$

其中,X_N和X_M表示中间特征;Y表示混合注意力Transformer模块的输出;将每个体素视为嵌入的token;MLP表示多层感知机。对于计算自注意力模块,给定形状为$H\times W\times L$的输入特征。在每个分辨率下,第一个混合注意力Transformer模块采用规则的窗口划分方法,从左上角的体素开始,将特征图均匀地划分为大小为$M_x\times M_y\times M_z$的非重叠窗口。然后在每个窗口内局部计算自注意力。基于窗口的自注意力被公式化为:

$$\mathrm{Attention}(Q,K,V)=\mathrm{SoftMax}(\frac{QK^T}{\sqrt{d}})V \tag{7}$$

其中,Q、K、$V\in R^{M_xM_yM_z\times d}$是来自单个图像的查询、键和值矩阵;$d$表示查询和键的维度。本文使用大尺寸的窗口来计算自注意力,因为大尺寸的窗口显著扩大所使用体素的范围。此外,为了在相邻的非重叠窗口之间建立连接,利用移位窗口划分方法,并将移位大小设置为窗口大小的一半。

CSA由两个标准卷积层组成,具有GELU激活函数[27]、CA以及空间注意力模块,如图3所示。由于基于Transformer的结构通常需要大量通道用于图像块的嵌入,因此直接使用具有恒定宽度的卷积会产生很大的计算成本。因此,本文使用常数b压缩两个卷积层的通道数。对于具有C个通道的输入特征,将第一个卷积层之后的输出特征的通道数压缩为C/b,然后通过第二层将该特征扩展为C个通道。接下来利用标准CA模块自适应地重新缩放信道特征。紧接着利用空间注意力对输入特征的不同位置进行加权处理,以便模型能够更有针对性地关注图片的边缘以及纹理等特定区域的细粒度语义信息。

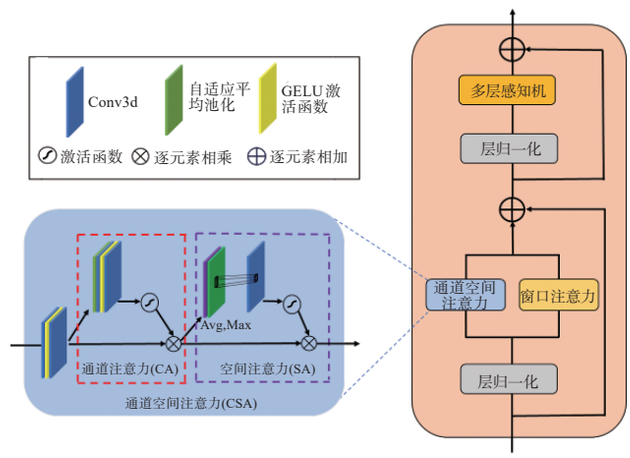

图3 混合注意力由两部分组成通道空间注意力和窗口注意力

Figure 3 Mixed attention consisting of channel spatial attention and window attention

1.4 损失函数

网络训练的总体损失函数源自传统图像配准算法的能量函数，损失函数由两部分组成：一部分计算变形的运动图像和固定图像之间的相似性，另一部分正则化变形场，使其平滑：

$$\hat{\phi} = \arg_{\phi}\min L(I_f, I_m, \phi) \tag{8}$$

$$L(I_f, I_m, \phi) = L_{\text{sim}}(I_f, I_m \circ \phi) + \lambda L_{\text{smooth}}(\phi) \tag{9}$$

其中，I_f表示固定图像；$I_m \circ \phi$表示运动图像I_m被变形场ϕ作用；函数$L_{\text{sim}}(\cdot,\cdot,\cdot)$测量$I_f$和$I_m \circ \phi$之间的图像相似性；$L_{\text{smooth}}(\cdot)$表示变形场正则化；$\lambda$是正则化参数。

图像相似性度量，L_{sim}使用的是局部归一化互相关：

$$\text{LNCC}(I_f, I_m, \phi) = \sum_{v \in \Omega} \frac{(\sum_{V_i}(I_f(V_i) - \bar{I}_f(V))([I_m \circ \phi])(V_i) - \left[\overline{I_m \circ \phi}\right](V))^2}{\left(\sum_{P_i}(I_f(V_i) - \bar{I}_f(V))^2\right)(\sum_{V_i}([I_m \circ \phi])(V_i) - \left[\overline{I_m \circ \phi}\right](V))^2} \tag{10}$$

其中，Ω表示图像大小；$\bar{I}_f$和$\overline{I_m \circ \phi}$分别表示固定图像$I_f$和扭曲图像$I_m \circ \phi$以体素$V$为中心的大小为$n^3$的局部窗口内的平均体素值，设$n$=9。

最小化相似性度量L_{smooth}以促进$I_m \circ \phi$在视觉上尽可能近似I_f。然而，由此产生的变形场ϕ可能不平滑或不真实。为了在变形场中施加光滑性，在损失函数中添加正则化惩罚项L_{smooth}。L_{smooth}促使一个位置中的位移值与其相邻位置中的值相似。在这里，使用的是扩散正则化[9]：

$$L_{\text{smooth}}(\phi) = \sum_{V \in \Omega} \left\| \nabla \phi(V) \right\|^2 \tag{11}$$

其中，$\phi(V)$是变形场ϕ的空间梯度，使用前向差来近似空间梯度，即：

$$\frac{\partial \phi(V)}{\partial \{x,y,z\}} \approx \phi(V_{\{x,y,z\}+1}) - \phi(V_{\{x,y,z\}}) \tag{12}$$

2 结果与分析

2.1 数据集与实验设置

本文在3个不同的大脑MRI数据集上评估所提方法的性能。对于LPBA40数据集[28]，每个MRI图片包含54个手动标记的感兴趣区域。LPBA40数据集中的所有体积都严格预对准MIN305（MNI305模板：MNI305是由305名健康成人大脑的平均结构生成的标准大脑模板，由蒙特利尔神经学研究所创建，它广泛用于脑部图像研究和分析。）。30幅图像（30×29对）用于训练，10幅图像（10×9对）用于测试。本文将图片裁剪为160×192×160，并使用专家标注的56个结构分割图作为评估的金标准。使用FreeSurfer[29]对每个MRI图片进行预处理，包括颅骨剥离、空间归一化、仿射变换和自动结构分割。OASIS数据集[30]从2021 Learn2Reg challenge获得，用于患者间配准。该数据集共包含451幅大脑T_1 MRI图像，其中94幅、19幅和38幅图像分别用于训练、验证和测试。使用FreeSurfer对大脑MRI图像进行预处理，并提供35个解剖结构的标签图进行评估。IXI数据集来自图像信息交换（IXI）数据库，它总共包含576幅T_1W脑部MRI图像，对其进行从标准图谱（Atlas）到患者的配准操作[31]。数据集分为403、58和115（7:1:2）幅图像，用于训练、验证和测试。类似地，对IXI数据集进行相同的预处理步骤。使用FreeSurfer对MRI图像进行预处理。将所有图像裁剪为160×192×224的大小。使用30个解剖结构的标签图来评估配准性能。

本文提出的网络在公共的PyTorch平台上实现。训练和测试使用CUDA的11.3版本，并在一个具有80 GB内存的NVIDIA A100 GPU上进行训练。对于LPBA40，所有模型都使用Adam优化进行1 000个时期的训练。批量大小为1。对于IXI和OASIS，所有模型都使用Adam优化进行500个时期的训练，批量大小为1，使用LNCC损失，初始学习率设为1×10^{-4}，权重衰减系数为0.001，训练轮次为500。

2.2 评价指标

为了定量评估配准性能，计算Dice相似性系数（DSC）[32]作为主要相似性度量，以评估相应区域之间的重叠程度，还提供这些指标的标准偏差（STD），以评估这些模型的稳定性。设$S_{I_f}^k$、$S_{I_{m\circ\phi}}^k$分别是I_f和I_m结构k上对应的体素，Dice相似系数为：

$$\text{Dice}(S_{I_f}^k, S_{I_{m\circ\phi}}^k) = 2 \cdot \frac{\left| S_{I_f}^k \bigcap (S_{I_{m\circ\phi}}^k) \right|}{\left| S_{I_f}^k \right| + \left| S_{I_{m\circ\phi}}^k \right|} \tag{13}$$

DSC越高表示配准精度越高。为了量化变形场的规律性，本文还列出雅可比矩阵行列式中非正值在变形场上的百分比（即$\left|J_\phi\right| \leqslant 0$），该度量越小，变形的纹理保持能力就越好。所有上述指标都是在3D中计算的。

2.3 对比实验

将本文方法与几种最先进的配准方法进行定性和定量的比较：（1）在SyN[33]中：使用均方差（MSQ）作为目标函数，同时使用默认的高斯平滑系数为3，以及分别进行180、80和40次3个不同尺度迭代。（2）在VoxelMorph[9]中，本文使用其官方的VoxelMoph-1实现，使用NCC损失以及L2平滑正则化项，此外正则化项的系数λ设置为1，学习率设置为0.000 4。（3）在CycleMorph[13]中，使用它的官方开源实现，将循环损失α的权重设置为0.1，单位损失β的权重设置0.5，平滑正则化项的权重λ设置为0.02，学习率设置为0.000 2。（4）在ViT-V-Net[34]中：该注册网络是基于ViT开发的，应用默认的网络超参数设置，本文使用均方

误差损失以及L2平滑正则化项，此外将正则化项的系数λ设置为0.02，以更好地拟合数据集，并将学习率设置为0.000 1。(5)在TransMorph[32]中，使用它的官方开源代码实现，使用NCC损失作为差异度量，正则化项的系数λ设置为1，学习率设置为0.000 1。

2.4 实验结果分析

在LPBA40数据集上，如表1所示，表中时间表示模型对齐一对图像所需的时间，对于传统方法，代表CPU计算时间，而对于基于深度学习的方法，则代表GPU计算时间。显存表示模型对齐一对图像所占用的显存大小。本文方法将DSC提高到0.665，优于其他4种基于深度学习的方法。本文提出方法的DSC比TransMorph高3.1%。此外，本文提出的网络在变形场中几乎没有折叠。与传统方法SyN相比，本文的网络在LPBA40数据集上的DSC值略低，但具有一定的速度优势。图4和图5为本文方法在LPBA40数据集上的可视化结果。图4表明本文方法生成更准确的配准图像，使用本文的方法可以始终如一地保留内部结构，并且具有更加细节的边缘以及纹理特征。图5显示本文方法生成更加平滑的变形场，网格折叠面积明显减少，具有明显的视觉优势。在IXI数据集上，如表1所示，在所有比较方法中，本文方法获得最佳的DSC值。本文的方法收敛速度的优势是显而易见的。与传统方法SyN相比，在DSC和速度方面都获得更好的性能。与4种深度学习方法相比，本文提出模型的DSC值比TransMorph高0.7%，并且DSC值最高。与其他4种深度学习方法相比，本文的方法实现更平滑的变形，并且几乎没有折叠。在OASIS数据集上，如表2所示，在所有比较方法中，本文的方法实现最高的DSC值和有竞争力的速度。

表1 不同模型在LPBA40和IXI数据集上的评价指标

Table 1 Evaluation indexes of different models on LPBA40 and IXI datasets

模型	LPBA40数据集				IXI数据集			
	DSC	$\%\lvert J_\phi\rvert \leqslant 0$	时间/s	显存/M	DSC	$\%\lvert J_\phi\rvert \leqslant 0$	时间/s	显存/M
SyN	0.698±0.017	<0.1	120.000	414	0.645±0.152	0.1	210.000	414
VoxelMorph	0.645±0.019	<0.7	0.229	10 366	0.729±0.129	1.5	0.201	11 775
CycleMorph	0.616±0.028	<0.6	0.192	16 478	0.737±0.123	1.7	0.195	24 219
ViT-V-Net	0.624±0.029	<0.2	0.165	17 330	0.734±0.124	1.6	0.319	17 388
TransMorph	0.634±0.033	<0.3	0.125	11 810	0.744±0.124	1.5	0.228	14 094
本文方法	0.665±0.014	<0.1	0.112	10 820	0.751±0.027	1.3	0.195	16 430

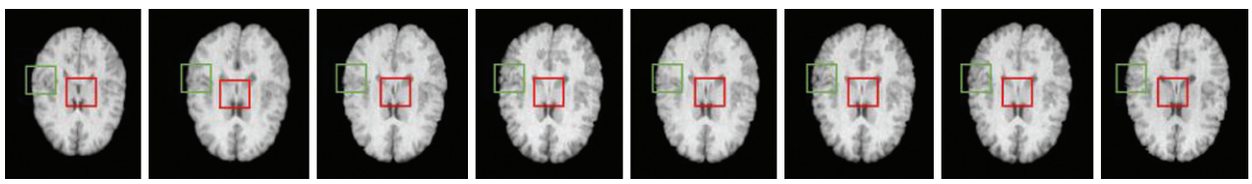

图4 不同模型在LPBA40数据集上的可视化效果对比

Figure 4 Comparison of visualizations of different models on LPBA40 dataset

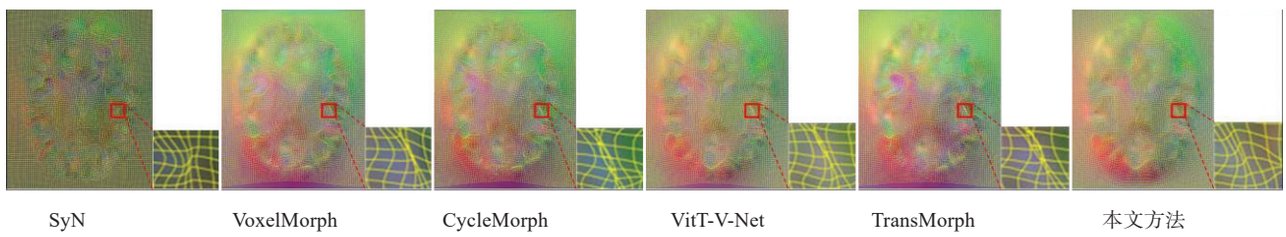

图5 不同模型在LPBA40数据集上生成的变形场网格的可视化图像

Figure 5 Visualization of the deformation field mesh generated by different models on LPBA40 dataset

表2 不同模型在OASIS数据集上的评价指标

Table 2 Evaluation indicators of different models on OASIS dataset

模型	DSC	$\%\lvert J_\phi\rvert \leq 0$	时间/s	显存/M
SyN	0.769±0.027	0.1	158.000	414
VoxelMorph	0.794±0.019	1.2	0.171	13 040
CycleMorph	0.803±0.023	1.2	0.529	22 876
ViT-V-Net	0.815±0.021	1.2	0.416	17 484
TransMorph	0.858±0.022	0.9	0.251	16 680
本文方法	0.860±0.013	0.5	0.249	16 518

2.5 消融实验

本节探究了网络中不同模块对于配准效果的影响,以证明模块的有效性。用模块一代表自适应上下文特征提取器,模块二混合注意力模块。如表3所示,与基线方法相比,本文方法实现了更平滑的变形,并且几乎没有折叠。在LPBA40数据集上,如表3所示,在所有比较方法中,本文方法实现最高的DSC值和有竞争力的速度。添加不同模块后各个指标均有不同程度的提升。以LPBA40数据集为例,添加模块一和模块二后,DSC分别提升1.2%和2.1%。同时添加模块一和二后体素折叠率和显存占用率明显下降,训练速度相较于原模型得到显著提升,表3的结果验证本文创新的有效性。

表3 提出的方法在LPBA40数据集上的消融实验

Table 3 Ablation study of the proposed method on LPBA40 dataset

模型	DSC	$\%\lvert J_\phi\rvert \leq 0$	时间/s	显存/M
基线	0.634±0.033	<0.3	0.125	11 810
添加模块一	0.646±0.037	<0.2	0.118	11 732
添加模块二	0.655±0.030	<0.2	0.105	10 764
添加模块一、二	0.665±0.014	<0.1	0.112	10 820

3 结 语

医学图像配准对于临床诊断和治疗具有非常重要的应用价值。本文将传统的特征提取方法与深度学习相结合,基于Transformer和CNN模型,提出一种无监督可变形图像配准框架。该模型采用SSC提取输入图像的体素邻域上下文语义信息,进行特征匹配,并提出一种新的混合注意力(STWA)去进一步开发Transformer的潜力。在3个3D大脑MRI扫描数据集上进行大量的图像配准实验,实验结果表明,与现有配准方法相比,实现更高的配准精度,验证本文方法实现了最先进的性能和配准效率,证明该方法在单模态医学图像配准任务中的有效性。在未来,将研究可以替代的损失函数,如互信息,以努力扩展所提出的方法在多模态医学图像配准任务中的潜力。

【参考文献】

[1] Sotiras A, Davatzikos C, Paragios N. Deformable medical image registration: a survey[J]. IEEE Trans Med Imaging, 2013, 32(7): 1153-1190.

[2] Wang HQ, Ni D, Wang Y. Recursive deformable pyramid network for unsupervised medical image registration[J]. IEEE Trans Med Imaging, 2024, 43(6): 2229-2240.

[3] Wang HQ, Wang ZY, Ni D, et al. ModeTv2: GPU-accelerated motion decomposition transformer for pairwise optimization in medical image registration[EB/OL]. (2024-03-25). https://arxiv.org/abs/2403.16526.

[4] Ma JY, Jiang XY, Fan AX, et al. Image matching from handcrafted to deep features: a survey[J]. Int J Comput Vis, 2021, 129(1): 23-79.

[5] Ma MR, Wang T, Song L, et al. RFR-WWANet: weighted window attention-based recovery feature resolution network for unsupervised image registration[EB/OL]. (2023-05-22). https://arxiv.org/abs/2305.04236v1.

[6] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16×16 words: transformers for image recognition at scale[EB/OL]. (2021-06-03). https://arxiv.org/abs/2010.11929.

[7] Wang HQ, Ni D, Wang Y. ModeT: learning deformable image registration *via* motion decomposition transformer[C]//Medical Image Computing and Computer Assisted Intervention-MICCAI 2023. Cham: Springer Nature Switzerland, 2023: 740-749.

[8] Chen ZY, Zheng YJ, Gee JC. TransMatch: a transformer-based multilevel dual-stream feature matching network for unsupervised deformable image registration[J]. IEEE Trans Med Imaging, 2024, 43(1): 15-27.

[9] Balakrishnan G, Zhao A, Sabuncu MR, et al. VoxelMorph: a learning framework for deformable medical image registration[J]. IEEE Trans Med Imaging, 2019, 38(8): 1788-1800.

[10] Onofrey JA, Staib LH, Papademetris X. Semi-supervised learning of nonrigid deformations for image registration[C]//Medical Computer Vision. Large Data in Medical Imaging. Cham: Springer, 2014: 13-23.

[11] Yang X, Kwitt R, Styner M, et al. Quicksilver: fast predictive image registration-a deep learning approach[J]. Neuroimage, 2017, 158: 378-396.

[12] Rohé MM, Datar M, Heimann T, et al. SVF-net: learning deformable image registration using shape matching[C]//Medical Image Computing and Computer Assisted Intervention-MICCAI 2017. Cham: Springer International Publishing, 2017: 266-274.

[13] Kim B, Kim DH, Park SH, et al. CycleMorph: cycle consistent unsupervised deformable image registration[J]. Med Image Anal, 2021, 71: 102036.

[14] Yuan W, Cheng J, Gong YH, et al. MACG-net: multi-axis cross gating network for deformable medical image registration[J]. Comput Biol Med, 2024, 178: 108673.

[15] Bahdanau D, Cho K, Bengio Y, et al. Neural machine translation by jointly learning to align and translate[EB/OL]. (2016-05-19). https://arxiv.org/abs/1409.0473.

[16] He KL, Gan C, Li ZY, et al. Transformers in medical image analysis [J]. Intell Med, 2023, 3(1): 59-78.

[17] Li J, Chen JY, Tang YC, et al. Transforming medical imaging with transformers? A comparative review of key properties, current progresses, and future perspectives[J]. Med Image Anal, 2023, 85: 102762.

[18] Liu QY, Kaul C, Anagnostopoulos C, et al. Optimizing vision transformers for medical image segmentation and few-shot domain adaptation[EB/OL]. (2022-10-26). https://arxiv. org/abs/2210. 08066v1.

[19] Heinrich MP, Jenkinson M, Papież BW, et al. Towards realtime multimodal fusion for image-guided interventions using self-similarities[C]//Medical Image Computing and Computer-Assisted Intervention-MICCAI 2013. Heidelberg: Springer, 2013: 187-194.

[20] Liu Z, Lin YT, Cao Y, et al. Swin transformer: hierarchical vision

transformer using shifted windows[C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway, NJ, USA: IEEE, 2021: 9992-10002.
[21] Dong XY, Bao JM, Chen DD, et al. CSWin transformer: a general vision transformer backbone with cross-shaped windows[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recogni tion (CVPR). Piscataway, NJ, USA: IEEE, 2022: 12114-12124.
[22] Raghu M, Unterthiner T, Kornblith S, et al. Do vision transformers see like convolutional neural networks? [C]//Proceedings of the 35th International Conference on Neural Information Processing Systems. Red Hook, NY, USA: Curran Associates Inc., 2024: 12116-12128.
[23] Heinrich MP, Jenkinson M, Bhushan M, et al. MIND: modality independent neighbourhood descriptor for multi-modal deformable registration[J]. Med Image Anal, 2012, 16(7): 1423-1435.
[24] Li KC, Wang YL, Zhang JH, et al. UniFormer: unifying convolution and self-attention for visual recognition[J]. IEEE Trans Pattern Anal Mach Intell, 2023, 45(10): 12581-12600.
[25] Zhao YC, Wang GT, Tang CX, et al. A battle of network structures: an empirical study of CNN, transformer, and MLP[EB/OL]. (2021-11-25). https://arxiv.org/abs/2108.13002v1.
[26] Wu HP, Xiao B, Codella N, et al. CvT: introducing convolutions to vision transformers[C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway, NJ, USA: IEEE, 2021: 22-31.
[27] Hendrycks D, Gimpel K. Gaussian error linear units (GELUs)[EB/OL]. (2023-06-06). https://arxiv.org/abs/1606.08415.
[28] Shattuck DW, Mirza M, Adisetiyo V, et al. Construction of a 3D probabilistic atlas of human cortical structures[J]. Neuroimage, 2008, 39(3): 1064-1080.
[29] Fischl B. FreeSurfer[J]. Neuroimage, 2012, 62(2): 774-781.
[30] Marcus DS, Wang TH, Parker J, et al. Open access series of imaging studies (OASIS): cross-sectional MRI data in young, middle aged, nondemented, and demented older adults[J]. J Cogn Neurosci, 2007, 19(9): 1498-1507.
[31] Chen JY, Frey EC, He YF, et al. TransMorph: transformer for unsupervised medical image registration[J]. Med Image Anal, 2022, 82: 102615.
[32] Dice LR. Measures of the amount of ecologic association between species[J]. Ecology, 1945, 26(3): 297-302.
[33] Avants BB, Epstein CL, Grossman M, et al. Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain[J]. Med Image Anal, 2008, 12(1): 26-41.
[34] Chen J, He Y, Frey E C, et al. Vit-v-net: vision transformer for unsupervised volumetric medical image registration[J]. arXiv preprint arXiv: 2104.06468, 2021.

（编辑：陈丽霞）