doi:10.3969/j.issn.1001-4616.2025.04.012

基于 Transformer 和 CNN 的双流网络眼底图像可变形配准方法

吴水淼,陈　强

(南京理工大学计算机科学与工程学院,江苏 南京 210094)

[**摘要**]　视网膜图像配准有助于医生全面了解视网膜结构,但由于缺乏带真实标签的数据,增加了配准难度.传统配准方法效率低,且依赖固定模型和手工设计特征,难以处理复杂变形,而深度学习方法尽管高效,但多为单流结构,该结构会对特征融合产生干扰. 针对现有方法在特征提取上的不足,本文提出一种双流网络,通过 Transformer 和 CNN 分别提取全局和局部特征,在多个尺度上进行特征匹配,并引入血管信息辅助训练. 实验结果表明,该方法在彩色眼底数据集上显著提升了配准精度,验证了其在可变形医学图像配准中的有效性.

[**关键词**]　眼底图像配准,无监督学习,transformer,CNN

[**中图分类号**]TP391　[**文献标志码**]A　[**文章编号**]1001-4616(2025)04-0118-10

Deformable Registration Method for Retinal Images Based on a Dual-Stream Network of Transformer and CNN

Wu Shuimiao, Chen Qiang

(School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China)

Abstract: Retinal image registration helps doctors to comprehensively understand the structure of the retina, but the lack of data with real labels increases the difficulty of registration. Traditional registration methods are inefficient and rely on fixed models and manually designed features, making it difficult to handle complex deformations, while deep learning methods, although efficient, are mostly single stream structures that can interfere with feature fusion. In response to the shortcomings of existing methods in feature extraction, this paper proposes a dual stream network that extracts global and local features through Transformer and CNN respectively, performs feature matching at multiple scales, and introduces vascular information to assist in training. The experimental results show that this method significantly improves the registration accuracy on color fundus datasets, verifying its effectiveness in deformable medical image registration.

Key words: retinal image registration, unsupervised learning, transformer, CNN

可变形图像配准是许多医学成像分析任务的基础. 它的作用是建立空间对应关系,以最大限度地减少一对图像中参考图像和移动图像之间的差异. 图像配准在彩色眼底图像上的应用具有重要的临床和研究意义,它涉及将来自不同时间点、不同视角或不同设备的眼底图像对齐,以便进行精准的比较和分析. 而人工检查两幅或多幅视网膜图像之间的可能的变化是费力、耗时且容易出错的. 可变形图像配准通常涉及两个步骤:用于全局对准的初始刚性配准,随后是具有更高自由度的非刚性变换. 通过对图像进行非刚性变形来实现精准对齐,处理图像中的形变、扭曲或弹性变化. 这种方法通过构建变形模型或优化变形参数,使得移动图像能够准确地与参考图像对齐. 通过图像配准,医生可以追踪眼底疾病的发展变化,如糖尿病视网膜病变(DR)、青光眼、老年黄斑变性(AMD)[1]等,评估治疗效果并决定采取适当的治疗方法,也可在多幅图像中整合信息以提高诊断准确性.

收稿日期:2024-09-24.

基金项目:国家自然科学基金项目(92370109、6217223)、中央高校基本科研基金项目(30921013105).

通讯作者:陈强,博士,教授,研究方向:图像分析与机器学习. E-mail:chen2qiang@ njust.edu.cn

为了避免使用特定的机制来生成标签数据,同时能够处理许多不同类别的眼底图像. 本文提出了一个端到端的无监督学习配准框架,网络在每幅图像上独立提取特征,为避免模型在特征融合过程中产生干扰,通过两个分支分别对需要配准的两幅图像进行特征提取,使得模型能够更准确地处理每幅图像的独特信息. 由于 Transformer[2] 在提取全局特征和处理长距离关系时具有天然良好的建模能力,其与 CNN 的归纳偏置特性具有较强互补性[3]. 本文将两者进行结合,前者负责捕获全局信息,利用自注意力机制处理图像中的长程依赖关系,帮助模型更好地理解眼底图像的整体结构;后者用于提取局部特征,特别是细节和纹理信息. 之后利用交叉注意力机制在多个尺度上进行特征匹配,逐级显式匹配不同分辨率下的特征图,确保特征在整个配准过程中能够精确对齐. 血管分割图像提供了眼底图像中的关键结构信息,使得配准算法能够更精确地对齐血管区域,从而提高配准的准确性.

1 相关工作

1.1 传统配准方法

基于优化的配准方法大致可以分为两类:基于强度的方法和基于特征的方法. 基于强度的方法通过比较图像的像素强度值来进行配准,而基于特征的方法则通过提取并匹配图像中的关键特征点来实现配准. 这些配准算法通常依赖于线性变换,例如刚性变换和仿射变换,旨在对两幅图像进行全局对齐. 而非刚性配准主要针对于两幅待配准图像间的局部对齐.

传统的非刚性配准方法一般遵循以下步骤:以 B 样条变形模型为例,它是 SimpleITK[4] 下的一个配准工具包,会为每对图像优化一个目标函数. 首先初始化模型参数,构建初始变形模型,并从待配准图像中提取特征. 随后,定义相似性度量以评估变形后图像之间的匹配程度,常见的度量指标有均方误差、归一化互相关等等,同时引入正则项以确保变形的合理性. 通过基于梯度的 L-BFGS-B[5] 优化算法迭代调整变形模型的参数,使目标函数达到最优解. 将优化后的变形模型应用于移动图像,生成配准后的图像. 最后,通过定量指标评估配准效果,确保配准精度达到预期要求. 然而,这样的方法在计算上是昂贵的,并且在实践中通常是缓慢的,因为优化问题需要针对每对看不见的图像重新求解,不适合在临床环境下应用.

1.2 深度学习配准方法

对于基于学习的配准方法,变形场和图像之间的相关性可以通过深度学习技术来学习. Gutierrez-Becker 等[6] 提出通过使用基于梯度优化的监督回归模型来估计变形参数,而不是直接最小化损失目标函数. 该方法证明了可以提高配准性能,但是该方法需要大规模的像素级标签. 然而,大规模的像素级标签是很难收集的,所以现有的主流方法主要采用无监督的方式.

基于无监督的配准方法通过优化图像之间的相似性度量来学习变形关系,而无需配准标签,通过图像间自我监督机制不断调整网络参数以提高配准精度. 大多数现有的可变形配准算法基于能量函数迭代地优化变换[7]. 令 $\boldsymbol{f}$ 和 $\boldsymbol{m}$ 分别表示参考图像和移动图像,并且令 $\boldsymbol{\phi}$ 是将 $\boldsymbol{m}$ 的坐标映射到 $\boldsymbol{f}$ 的坐标的配准域. 优化问题一般可以描述为:

$$\boldsymbol{\phi}=\arg\min_{\phi}\mathcal{L}(\boldsymbol{f},\boldsymbol{m},\boldsymbol{\phi}), \tag{1}$$

$$\mathcal{L}(\boldsymbol{f},\boldsymbol{m},\boldsymbol{\phi})=\mathcal{L}_{\text{sim}}(\boldsymbol{f},\boldsymbol{m}\circ\boldsymbol{\phi})+\lambda\mathcal{L}_{\text{smooth}}(\boldsymbol{\phi}), \tag{2}$$

式中,$\boldsymbol{m}\circ\boldsymbol{\phi}$ 表示移动图像被变形场扭曲,$\mathcal{L}_{\text{sim}}(\cdot,\cdot)$ 用于度量两幅图像的相似性,$\mathcal{L}_{\text{smooth}}(\cdot)$ 表示对变形场添加梯度约束,帮助限制变形场的剧烈变化,确保变形场在空间上是平滑的,λ 是正则化权衡参数.

与传统方法相比,基于深度学习的方法可以显著提高配准性能,同时操作速度更快. 近几年,Balakrishnan 等[8] 提出了一种基于深度学习的医学图像配准框架 VoxelMorph,将移动图像和参考图像在通道上拼接,利用 CNN 来学习图像之间的变形,允许网络在配准过程中自动学习特征和变形场,通过 STN[9] 网络来对移动图像进行变形,展示了深度学习在图像配准中的潜力. 这种方法减少了手动特征提取的需求,并提高了模型的适应性,与传统的基于优化的配准方法相比,能够显著提高配准的速度. Chen 等[10] 提出了 TransMorph 模型,一种结合 Transformer 与 ConvNet 的混合模型,用于脑类数据图像配准,发现 ConvNets 在捕捉图像中的远程空间关系方面存在局限性,提出利用 Swin-Transformer[11] 在特征提取方面具有较大感受野的优势,采用分层结构和局部窗口自注意机制,能够捕捉不同尺度的特征,通过窗口移动策略增强跨窗口信息交流,提高特征提取的全面性并且效率更高,在图像配准任务中可以更精确地理解移动图像和参考

图像之间的空间对应关系. Chen 等[12]提出了一种新的基于 Transformer-CNN 的端到端双流无监督框架 TransMatch,在编码阶段将每幅图像通过独立分支利用 Swin-Transformer 进行特征提取分层特征,并结合 Transformer 的交叉注意机制对提取的特征实现多级显式特征匹配,最后通过 CNN 解码器输出形变场. 然而,现有方法往往依赖于单一的特征提取方式,如仅使用 CNN 或 Transformer,未能充分结合两者的优势. 这导致在处理复杂的解剖结构时,存在局部特征提取不足和鲁棒性差的问题. 本文的主要贡献如下:

(1)设计了一种新的双流网络架构,通过两个独立分支分别处理参考图像和移动图像,确保特征提取的独立性和完整性. 为克服单一特征提取方式的局限,编码阶段同时使用 CNN 和 Transformer 分别提取局部和全局特征,从而提高模型的综合特征表示能力.

(2)将眼底图像的血管信息作为辅助特征融入模型训练,提供额外的几何和解剖结构信息,提高模型的特征提取能力和配准精度,使得模型在处理复杂的眼底图像时更加准确和可靠.

(3)在公开的彩色眼底数据集上,我们的方法展现了更优异的整体表现. 实验结果表明,所提出的模型在配准精度和效率方面均优于现有的方法,验证了我们创新方法的有效性和实用性.

2 双流网络眼底图像可变形配准方法

2.1 刚性配准

本文采用 LoFTR[13]预训练模型来对眼底图像进行刚性配准,保证图像整体结构大致对齐. LoFTR 与传统刚性方法依赖于特征检测、描述和匹配的分阶段流程不同,它采用无检测器的方式,直接在像素级别上进行匹配. 核心思想是利用 Transformer 架构中的自注意力和交叉注意力机制来处理从卷积神经网络提取的密集特征,建立特征之间的匹配关系. 能够有效应对低纹理、大视角和光照变化等挑战,提供更准确的匹配结果.

2.2 可变形配准

我们提出了一种新颖的图像配准方法,通过结合双流网络结构、ResNet[14]与 Swin-Transformer,以及交叉注意力模块各自的优势,旨在提升图像配准的精度与鲁棒性. 具体而言,我们设计了两个独立的分支网络,分别处理参考图像和移动图像,保证了特征提取过程的独立性. 编码阶段,我们采用预训练的 ResNet-18 模型和 Swin-Transformer 模型分别为每幅图像提取局部和全局特征,将提取到的分层特征在通道上拼接融合后,使用交叉注意力模块对不同层次的特征进行匹配. 解码阶段,通过跳跃连接和卷积上采样操作,融合并恢复高分辨率特征,最终输出变形场. 另外对眼底图像进行血管分割,并将血管信息作为辅助特征融入配准训练,它可以提供额外的几何信息和解剖学结构,有助于提高模型的特征提取能力. 这一额外信息帮助模型更好地对齐血管区域,将模型预测的变形场作用于移动图像的分割结果上,通过对比变形后的血管分割图和参考图像的血管分割图,保证变形后解剖结构的一致性,利用这些信息可以增强模型对局部和全局特征的捕捉能力,从而提高配准的整体精度. 图 1 为网络模型架构概览图.

2.2.1 ResNet 模块

考虑到模型计算效率,该网络模块使用 ResNet-18 预训练权重对该模块进行初始化,由于 ResNet-18 在大规模自然数据集如 ImageNet 上进行了预训练,已经学习到了丰富的特征表示,这些特征表示在很多不同的计算机视觉任务中都表现出良好的泛化能力. 通过使用预训练模型,可以避免从零开始训练网络,从而加速训练过程并提高模型的性能. 并且该模块通过多个残差模块,能够提取到丰富且多样的特征表示. 对于图像配准任务,这种丰富的局部特征表示有助于捕捉图像之间的细微差异,提高配准的精度. 模型架构图如图 2 所示.

2.2.2 自注意力模块

Transformer 自注意力机制相比 CNN 能够在特征提取阶段捕捉长距离依赖和全局上下文信息,使得模型在处理多种视网膜场景和细节时更具鲁棒性. 对于每个尺度的特征图,分别计算查询 $\boldsymbol{Q}$、键 $\boldsymbol{K}$ 和值 $\boldsymbol{V}$,通过点积计算 $\boldsymbol{Q}$ 和 $\boldsymbol{K}$ 之间的相似度,得到注意力权重矩阵. 注意力权重反映了每个位置的特征对其他位置的特征的重要性. 接着,将这些权重应用于 $\boldsymbol{V}$ 向量,计算加权和,生成新的图像块表示. 自注意力计算过程如图 3 所示.

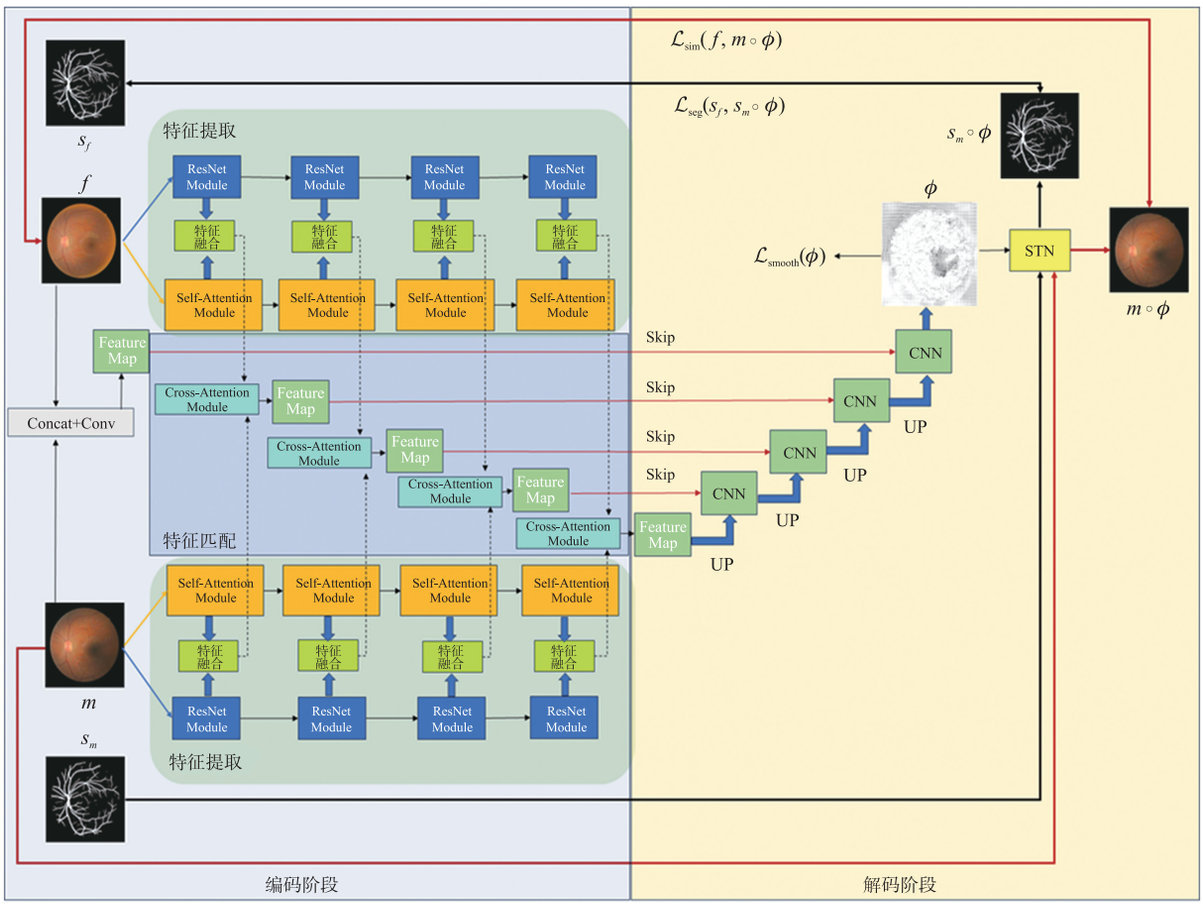


图 1　网络模型架构概览

Fig. 1　Overview of network model architecture

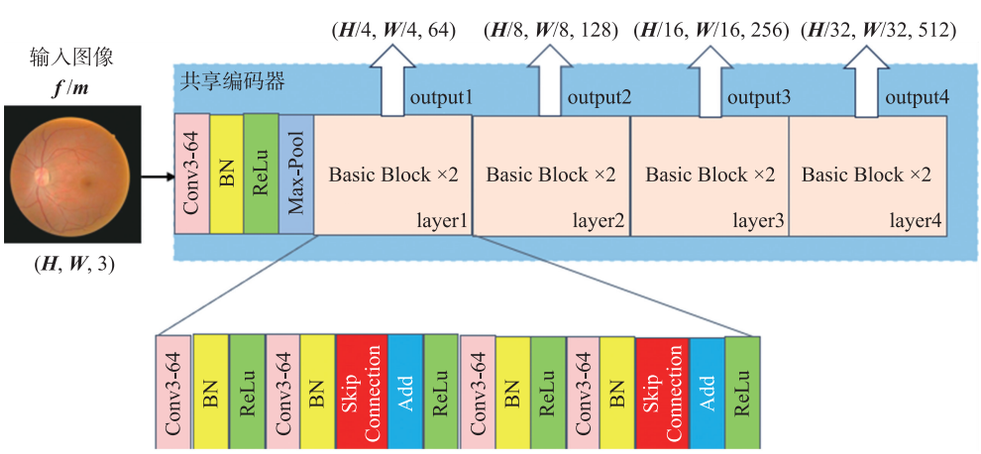


图 2　ResNet-18 模型架构图

Fig. 2　ResNet-18 model architecture diagram

2.2.3　交叉注意力模块

首先,从图像 $\boldsymbol{f}$ 中生成查询 $\boldsymbol{Q}$,而从图像 $\boldsymbol{m}$ 中生成键 $\boldsymbol{K}$ 和值 $\boldsymbol{V}$. 通过计算查询与键之间的相似度,使用 softmax 函数将相似度转换为权重,这些权重用于加权相应的值. 最后,将加权后的值进行求和,得到交叉注意力的输出. 这一过程使得图像 $\boldsymbol{f}$ 的特征能够与图像 $\boldsymbol{m}$ 的特征进行有效的匹配和融合,从而提升图像配准的精度和效果. 交叉注意力计算过程如图 4 所示.

2.2.4　空间变换网络

为了使网络预测出的变形场作用到移动图像进行变形,并且能够使用标准的基于梯度的方法来完成网络的误差反向传播,通过引入空间变换功能来增强网络对输入图像的处理能力. STN[9] 模块由两个主要

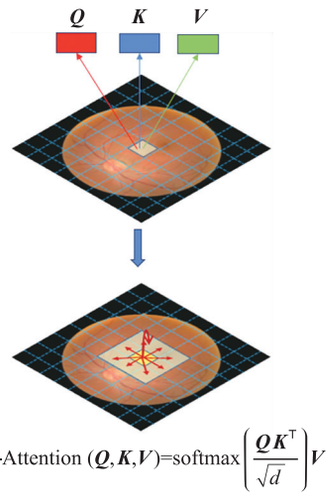


图 3　自注意力计算

Fig. 3　Self-attention computation

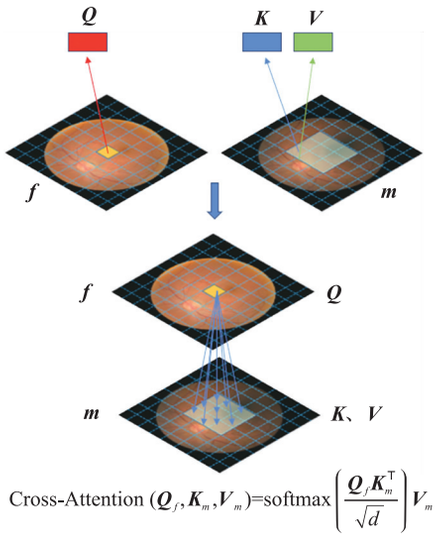


图 4　交叉注意力计算

Fig. 4　Cross-attention computation

部分组成:一个采样网格生成器和一个可微分的图像采样器. 对于移动图像中每个像素,计算该像素处的位移:$\boldsymbol{p}'=\boldsymbol{p}+\boldsymbol{u}(\boldsymbol{p})$,由于图像像素位置坐标只允许整数值,因此在 4 个相邻像素处进行双线性插值:

$$\boldsymbol{m}\circ\boldsymbol{\phi}(\boldsymbol{p})=\sum_{\boldsymbol{q}\in\boldsymbol{Z}(\boldsymbol{p}')}\boldsymbol{m}(\boldsymbol{q})\sum_{d\in x,y}(1-|\boldsymbol{p}'_d-\boldsymbol{q}_d|) \tag{3}$$

式中,$Z(\boldsymbol{p}')$表示 $\boldsymbol{p}'$的相邻像素,d 表示变形图像中每个像素位置.

2.2.5　损失函数

模型损失函数$\mathcal{L}(\boldsymbol{f},\boldsymbol{m},\boldsymbol{s}_f,\boldsymbol{s}_m,\boldsymbol{\phi})$由 3 个分量组成:惩罚外观差异的$\mathcal{L}_{\text{sim}}(\cdot,\cdot)$、惩罚 $\boldsymbol{\phi}$ 中的局部空间变化的$\mathcal{L}_{\text{smooth}}(\cdot)$和用于辅助训练的$\mathcal{L}_{\text{seg}}(\cdot,\cdot)$:

$$\mathcal{L}(\boldsymbol{f},\boldsymbol{m},\boldsymbol{s}_f,\boldsymbol{s}_m,\boldsymbol{\phi})=\mathcal{L}_{\text{sim}}(\boldsymbol{f},\boldsymbol{m}\circ\boldsymbol{\phi})+\lambda\,\mathcal{L}_{\text{smooth}}(\boldsymbol{\phi})+\gamma\,\mathcal{L}_{\text{seg}}(\boldsymbol{s}_f,\boldsymbol{s}_m\circ\boldsymbol{\phi}). \tag{4}$$

第一项是相似性损失采用 MSE,当 $\boldsymbol{f}$ 和 $\boldsymbol{m}$ 具有相似的图像强度分布和局部对比度时适用:

$$\mathbf{MSE}(\boldsymbol{f},\boldsymbol{m}\circ\boldsymbol{\phi})=\frac{1}{|\Omega|}\sum_{(x,y)\in\Omega}[\boldsymbol{f}(x,y)-[\boldsymbol{m}\circ\boldsymbol{\phi}](x,y)]^2. \tag{5}$$

最小化$\mathcal{L}_{\text{sim}}$将促使 $\boldsymbol{m}\circ\boldsymbol{\phi}$ 逼近 $\boldsymbol{f}$,但可能会产生物理上不真实的非平滑的 $\boldsymbol{\phi}$. 因此我们在位移 $\boldsymbol{u}$ 的空间梯度上使用扩散正则化器来鼓励平滑位移场 $\boldsymbol{\phi}$:

$$\mathcal{L}_{\text{smooth}}(\boldsymbol{\phi})=\sum_{\boldsymbol{p}\in\Omega}\|\nabla\boldsymbol{u}(\boldsymbol{p})\|^2. \tag{6}$$

特别地,$\nabla\boldsymbol{u}(\boldsymbol{p})=\left(\frac{\partial\boldsymbol{u}(\boldsymbol{p})}{\partial\boldsymbol{x}},\frac{\partial\boldsymbol{u}(\boldsymbol{p})}{\partial\boldsymbol{y}}\right)$,其中$\frac{\partial\boldsymbol{u}(\boldsymbol{p})}{\partial\boldsymbol{x}}\approx\boldsymbol{u}((p_x+1,p_y))-\boldsymbol{u}((p_x,p_y))$,$\frac{\partial\boldsymbol{u}(\boldsymbol{p})}{\partial\boldsymbol{y}}$也进行相同近似计算. 最后一项是用于辅助训练的$\mathcal{L}_{\text{seg}}$,设 $\boldsymbol{s}_f$ 和 $\boldsymbol{s}_m\circ\boldsymbol{\phi}$ 分别表示参考图像和经过变形的移动图像的血管分布图像,通过 DICE[15] 得分来衡量血管的对齐情况:

$$\text{DICE}(\boldsymbol{s}_f,\boldsymbol{s}_m\circ\boldsymbol{\phi})=2\cdot\frac{|\boldsymbol{s}_f\cap(\boldsymbol{s}_m\circ\boldsymbol{\phi})|}{|\boldsymbol{s}_f|+|\boldsymbol{s}_m\circ\boldsymbol{\phi}|}. \tag{7}$$

DICE 值为 1 表示经过变形后血管完全对齐,值为 0 表示血管没有重叠区域,血管分割损失定义如下:

$$\mathcal{L}_{\text{seg}}(\boldsymbol{s}_f,\boldsymbol{s}_m\circ\boldsymbol{\phi})=-\text{DICE}(\boldsymbol{s}_f,\boldsymbol{s}_m\circ\boldsymbol{\phi}). \tag{8}$$

3　实验

3.1　数据集

实验采用公共的眼底图像配准数据集 FIRE[16],该数据集由 129 幅视网膜图像组成,形成 134 对图像. 这些图像对根据其特征分为 3 个不同的类别,分辨率大小为 2 912×2 912 px,它们各自的特征和数量如表 1 所示。

表 1　FIRE 数据集各类别数据情况

Table 1　Data of each category of the FIRE dataset

	重叠程度	结构特征	数量/对
A	高	含病变	14
P	低	含黑色团块状	49
S	高	血管结构清晰	71

3.2 预处理

为了保证模型能够在不同状况下完成对眼底图像准确对齐,一些预处理操作是有必要的. 首先,对原始数据使用 LoFTR 进行刚性配准,保证整体结构对齐. 另外,为减少计算资源的消耗,将图像缩放到 576×576 px.

其次,由于眼底图像存在照度不均、噪声干扰等现象,直接对彩色眼底图像进行血管分割的效果不理想. 因此,限制对比度的自适应直方图均衡化(CLAHE)常被应用于视网膜图像预处理. 该步骤参考吴等[17]对视网膜图像的预处理方式,提取彩色图像中的绿色通道,归一化后利用 CLAHE 增强血管与背景之间的对比度,使用对比度增强来凸显血管和背景的强度差异.

此外为辅助下一步可变形配准模型的训练,使用 UNet[18]预训练模型对数据进行血管分割,能够更好地引导模型优化配准效果,确保模型关注图像中最重要的特征区域,对比度增强及血管分割结果图 5 所示.

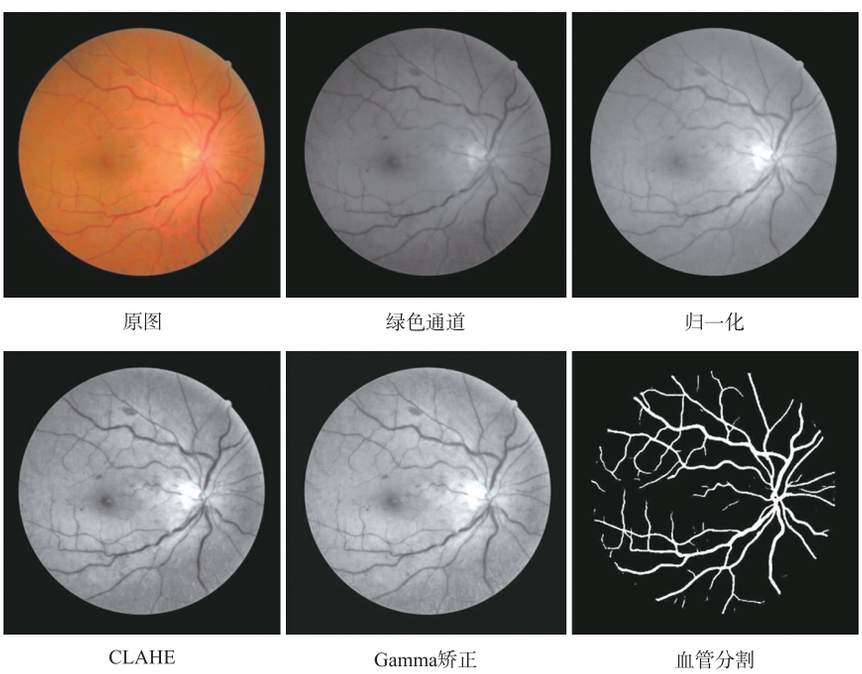


图 5 对比度增强及血管分割结果

Fig. 5 Contrast enhancement and vascular segmentation results

最后,为使得模型可以学习到更多样化的特征,数据集划分为 54 对训练集和 80 对测试集,并对训练集进行数据增广,通过仿射变换扩充为 108 对,针对每对图像应用相同的小范围的单应性变换. 具体而言,对于每对图像,对参考图像和移动图像随机应用以下变换之一:平移、缩放、旋转或剪切,对应血管分割图也采用相同的单应性变换,从而使模型在面对不同的图像变形时表现得更为鲁棒,减少对模型的过拟合的影响.

3.3 实验设置

我们的方法是基于 Pytorch 框架实现,依据现有配准任务研究的经验,将初始学习率设为 1e-4,相似性损失采用 MSE,λ 取 0.5,γ 取 1.0,当 λ 取值过大时,会过度强调变形场的平滑性,限制图像的变形能力,取值过小时,会出现不合理的形变;γ 值过大时,会过度优化针对分割图像的对齐,忽视图像本身,取值过小会对重要结构区域缺乏约束,造成关键解剖结构的配准精度下降;由于训练过程中,每对输入图像都有对应的分割标签,需保证每个 batch 中原始图像和分割图像通道数大小相同,所以 batch 取为 1,训练轮次为 5 000,所有实验均在配备有 Nvidia 3060Ti 的系统上进行,为性能测量提供一致的硬件标准. 并采用具有学习率衰减策略的 Adam[19]优化器,保证训练模型能够收敛到最优解,学习率更新策略如下:

$$lr_n = lr_{\text{init}} \cdot \left(1 - \frac{n-1}{N}\right)^{0.9},\ n = 1, 2, \cdots, N \tag{9}$$

式中,lr_n 表示第 n 轮的学习率,lr_{init} 表示初始学习率. 在图像配准任务中,模型需要学习到最优的变形场来对齐图像,这种逐步递减的学习率衰减策略能够在初期使用较大的学习率快速逼近最优解区域,而在后期逐渐减小学习率,从而更细致地优化模型参数,避免过拟合和欠拟合的情况,并且有助于模型在图像配准

任务中更精确地捕捉图像之间的几何关系.

3.4　对比实验

将我们的模型与先前已证明具有最先进配准性能的各种配准方法进行了比较:(1)Bspline:一种传统的迭代图像配准技术,通过调整控制点来定义变形场,以实现高精度的图像对齐. (2)VoxelMorph:一种基于 CNN 单流配准网络. (3)TransMorph:一种编码采用 Swin-Transformer 增强特征提取,解码阶段采用 CNN 的单流配准网络. (4)TransMatch:一种通过自注意力增强特征提取并通过交叉注意力来实现特征匹配的双流配准网络. 对比实验超参数都是采用跟本方法一致的设置方案,以平衡配准精度和运行时间之间的权衡.

3.5　消融实验

为评估在编码阶段的不同位置添加 ResNet 模块对模型性能的影响,从而验证在编码阶段各层次增加 CNN 提取特征的有效性. 通过逐层依次添加 ResNet 模块进行实验,每次实验仅在编码阶段的特定层次添加 ResNet 模块,其他层次仅使用 Swin-Transformer 来提取特征. 通过消融实验,验证结合 CNN 和 Transformer 特征的有效性,并为进一步优化模型结构提供理论依据.

为评估血管分割信息在图像配准中的辅助作用是否有效. 通过消融实验,在我们的方法上,一个是将血管分割图像与原始图像一起输入模型进行辅助训练,另一个则仅使用原始图像进行训练,其他实验设置保持一致. 通过对比这两种设置下的模型性能,旨在验证血管分割信息对提升配准精度的有效性,并分析其在配准过程中的具体作用机制.

3.6　评价指标

为了全面评估配准模型的性能,我们选择了 DICE、SSIM[20]、Time 作为主要评价指标. 需要注意的是,评价仅针对配准后的变形图像和参考图像的重叠区域进行,而不是整幅图像,这种方法的优点是能够更加准确地反映配准模型的配准性能,避免了无意义区域对评价结果造成干扰. 另外将模型预测出的变形场作用到分割图像上来验证配准的精度.

DICE 系数直观反映了配准后血管分割图与参考分割图的重合程度. SSIM 能够评估血管变形后的结构一致性,高 SSIM 值表示变形结果的结构保真度较高,而低 SSIM 值则指示变形结果可能存在结构失真或边缘模糊问题. Time 衡量配准的效率高低.

3.7　结果分析

在图 6 中,展示了不同配准方法定性分析结果,LoFTR 方法表示刚性配准结果,其他方法均为可变形配准方法,图中红色线条表示来自参考图像的血管,绿色线条表示来自变形图像的血管,黄色线条表示对齐的

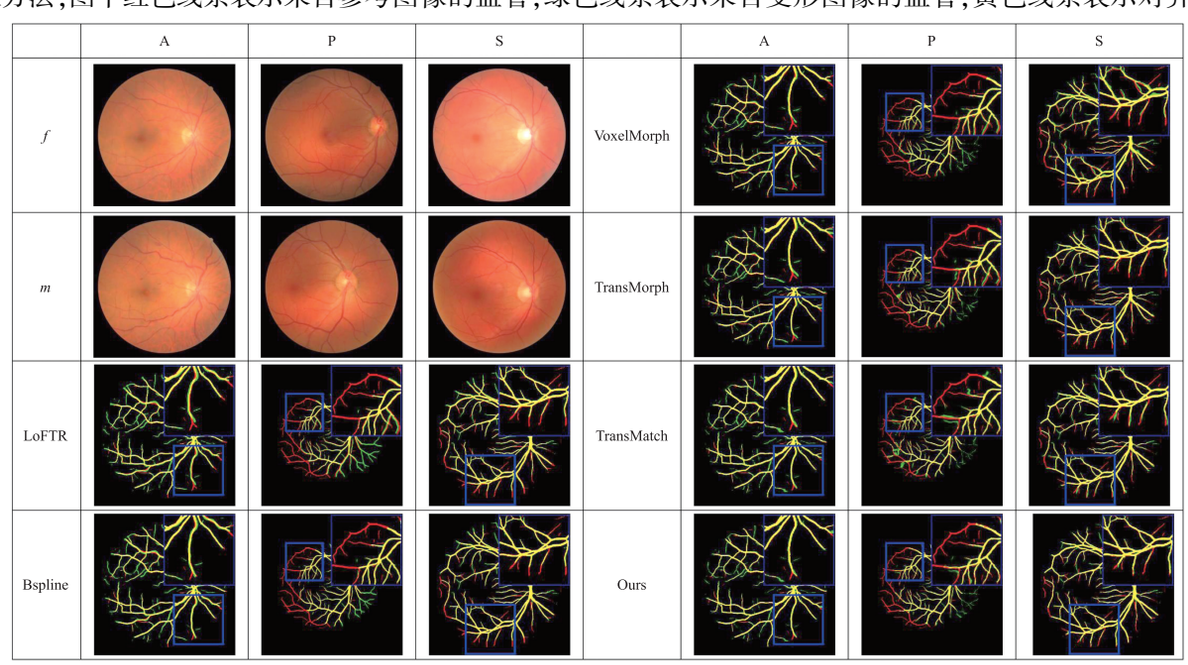


图 6　不同配准方法定性分析

Fig. 6　Qualitative analysis of different registration methods

血管,图片右上角表示局部区域放大后的配准结果. 实验结果显示,由于 LoFTR 不涉及非刚性变形,复杂血管结构的错配问题较为明显,这一结果也突出了后续可变形配准方法在复杂形变处理上的必要性和优越性;Bspline 在全局的血管配准效果较为理想,但在边缘复杂血管和局部区域依然存在错配,表明 Bspline 在处理局部细节丰富区域时的局限性;VoxelMorph 只利用卷积进行特征提取和变形场预测,配准效果一般,边缘区域未对齐血管较多;TransMorph 通过结合 Transformer 的全局特征提取能力和 CNN 的局部特征恢复能力,整体配准尚可,与 Voxelmorph 相比配准效果略好,但局部细节有误差,在放大区域中,红绿线条的错配现象依然存在,表明单流架构在复杂特征匹配中的局限性;TransMatch 采用双流架构,在编码阶段利用 Transformer 分别对参考图像和移动图像进行特征提取,并结合交叉注意力进行特征匹配,配准效果接近理想,血管配准中表现出较高的准确性,尽管如此,边缘细节错配现象依然存在,红绿线条仍有部分未能完全对齐,表明在边缘局部细节处理上仍有进一步提升的空间;我们的方法在 TransMatch 基础上,在特征提取阶段引入 ResNet 模块,与 Swin-Transformer 相互结合,共同提取更加丰富的局部和全局特征,与其他方法相比,配准效果最好,黄色重叠区域最多,特别是边缘复杂区域,我们的方法展现出更好的细节处理能力.

在表 2 中,LoFTR 方法表示刚性配准后的实验结果,其他方法均为可变形配准. 结果显示,我们的方法在 DICE 和 SSIM 指标上总体表现优秀,与其他方法相比,特别是在 A 类和 S 类的 DICE 上表现最好,分别达到了 0.719 1 和 0.792 4,并在 SSIM 指标上也有强劲的表现,达到了 0.650 6,仅次于 Bspline. 对于 P 类重叠度低的图像上,我们的方法在 DICE 指标上相比其他深度学习方法效果更好,但相比传统方法仍有差距. 但是在配准效率上,我们方法表现相比 Bspline 仍具有很大优势,能够在提高配准效率的前提下,提升配准精度.

在表 3 中,我们通过在编码阶段的不同层次添加 ResNet 模块来分析其对模型性能的影响,具体改进是在特定的特征提取层次上对原先只采用 Swin-Transformer 模块提取特征进行修改,额外增加 ResNet 模块来提取局部特征,如表中第一层则表示将 ResNet 的第一层输出特征($H/4$,$W/4$,64)与 Swin-Transformer 对应层次的输出特征($H/4$,$W/4$,96)在通道上进行拼接,输入到交叉注意力模块中,ResNet 模块和 Swin-Transformer 模块在各个层次的输出特征都是比前一层的输出图像通道数翻倍,尺寸减半. 依此类推,在其他

表 2　不同配准方法定量分析

Table 2　Quantitative analysis of the different registration methods

指标	方法	A	P	S
DICE(↑)	LoFTR[13]	0.629 6	0.572 8	0.772 3
	Bspline[4]	0.704 6	**0.667 0**	0.784 9
	VoxelMorph[8]	0.677 1	0.628 8	0.768 2
	TransMorph[10]	0.681 4	0.616 8	0.777 8
	TransMatch[12]	0.712 8	0.630 1	0.791 9
	Ours	**0.719 1**	0.650 6	**0.792 4**
SSIM(↑)	LoFTR	0.769 8	0.917 8	0.833 5
	Bspline	**0.817 3**	**0.937 5**	**0.865 1**
	VoxelMorph	0.798 4	0.924 5	0.842 0
	TransMorph	0.798 4	0.917 7	0.849 2
	TransMatch	0.813 1	0.911 5	0.855 2
	Ours	0.817 1	0.921 9	0.856 4
Time/s(↓)	Bspline	27.061 6	44.968 0	22.188 7
	VoxelMorph	**0.007 0**	**0.006 6**	**0.006 5**
	TransMorph	0.041 2	0.040 9	0.041 0
	TransMatch	0.099 3	0.097 7	0.097 7
	Ours	0.149 7	0.148 6	0.148 0

表 3　ResNet 模块添加位置消融实验结果

Table 3　Ablation study results for ResNet module addition positions

指标	添加位置	A	P	S
DICE(↑)	第一层	0.728 1	**0.652 8**	0.796 9
	第二层	0.713 6	0.636 6	0.789 2
	第三层	0.697 5	0.610 2	0.783 2
	第四层	0.693 5	0.605 6	0.780 7
	第一、二层	**0.729 1**	0.648 2	**0.800 7**
	第一、二、三层	0.717 2	0.631 4	0.791 7
	所有层	0.719 1	0.650 6	0.792 4
SSIM(↑)	第一层	**0.822 0**	0.921 0	0.859 6
	第二层	0.812 3	0.916 9	0.853 5
	第三层	0.803 2	0.907 9	0.848 0
	第四层	0.804 0	0.906 2	0.848 0
	第一、二层	0.821 7	0.916 3	**0.860 8**
	第一、二、三层	0.815 0	0.916 4	0.855 7
	所有层	0.817 1	**0.921 9**	0.856 4
Time/s(↓)	第一层	0.171 7	0.121 6	0.100 0
	第二层	0.144 2	0.108 1	0.102 0
	第三层	0.140 3	0.091 5	0.087 6
	第四层	0.150 3	0.116 0	0.103 7
	第一、二层	0.141 4	**0.090 9**	0.088 3
	第一、二、三层	**0.138 9**	0.091 8	**0.086 6**
	所有层	0.149 7	0.148 6	0.148 0

层次上的修改也是将两种模块的输出特征执行相同的操作,其他网络架构与 TransMatch 模型保持一致. 实验结果显示,在前两层添加该模块时,模型的 DICE 和 SSIM 均有所提升,表明在这些层次上,ResNet 模块能够有效地增强局部特征提取能力,从而提高模型的配准性能,因为前两层通常负责提取低级别的特征,包括图像中的边缘、纹理等,ResNet 模块擅长捕捉局部特征,能够有效增强这些层次的特征表示,为高层特征提供更可靠的基础,从而提高模型在捕捉细节上的表现. 而在后两层添加 ResNet 模块表现并未带来提升,引入了额外的复杂性,因为后两层通常负责提取更高级别的特征,特征通常是全局信息,不再局限于局部区域,此时的特征是经过多层抽象后的综合特征,模型更加依赖全局特征,ResNet 的局部增强作用变得相对有限. 并且前两层添加 ResNet 模块不仅能提升精度,还能保持较高的运行效率,而过多层次的 ResNet 模块并不会带来进一步的性能提升,反而可能增加计算复杂度. 综上所述,适度的局部增强是有益的,但过度使用可能带来负面影响,我们明确得出在编码阶段的前两层添加 ResNet 模块是最优选择,能够在提升模型配准精度的同时,保持较高的配准效率.

在表 4 中,展示了是否增加血管信息对模型性能影响的消融实验结果,由于血管信息的引入不会增加模型的参数复杂度,因此主要讨论其对配准精度的影响. 实验结果显示,不添加血管信息时,三类数据的 DICE 和 SSIM 指标均显著下降,尤其是在 S 类数据中,DICE 从 0.792 4 下降到 0.185 5,SSIM 从 0.856 4 下降到 0.604 9. 这表明在没有血管信息的情况下,模型的配准精度大幅降低. 主要原因在于缺乏明确的结构引导,导致模型在复杂配准任务中容易发生错误匹配,并且在优化过程中忽视关键解剖结构,最终导致配准精度下降. 而通过引入血管信息,可以增强模型对关键结构的约束,确保模型在训练时更稳定地朝着正确方向收敛,减少误配现象. 消融实验进一步证明了血管信息在眼底图像配准任务中的重要性,不仅提升了配准精度,还确保了关键解剖结构的准确对齐,验证了其作为辅助特征的必要性. 总之,消融实验结果明确表明,血管信息的引入对于实现高精度和稳定的眼底图像配准至关重要.

表 4 添加血管信息消融实验结果

Table 4 Ablation study results for adding vascular information

指标	是否添加	A	P	S	指标	是否添加	A	P	S
DICE	否	0.148 6	0.137 0	0.185 5	SSIM	否	0.603 7	0.812 8	0.604 9
(↑)	是	**0.719 1**	**0.650 6**	**0.792 4**	(↑)	是	**0.817 1**	**0.921 9**	**0.856 4**

4 结论

本文提出了一种基于 Transformer-CNN 的双流网络眼底图像可变形配准方法,利用 Transformer 和 CNN 模型的互补性,增强特征提取能力,配合交叉注意力模块对分层特征进行逐级匹配. 采用结构驱动的方式,通过血管分割变形结果辅助网络训练,减少模型在血管区域的配准误差,进而提高配准精度. 实验结果显示,我们的配准方法在 A、P、S 这三类数据上的表现均优于其他方法,特别是在血管细节和重叠区域的处理上更为准确和精细,这表明我们的方法在配准精度和细节处理方面具有显著优势.

[参考文献]

[1] LIM L S,MITCHELL P,SEDDON J M,et al. Age-related macular degeneration[J]. The lancet,2012,379(9827):1728-1738.

[2] VASWANI A,SHAZEER N,PARMAR N,et al. Attention is all you need[J]. Advances in neural information processing systems,2017,30:5998-6008.

[3] 石磊,籍庆余,陈清威,等. 视觉 Transformer 在医学图像分析中的应用研究综述[J]. 计算机工程与应用,2023,59(8):41-55. DOI:10.3778/j.issn.1002-8331.2206-0022.

[4] LOWEKAMP B C,CHEN D T,IBÁÑEZ L,et al. The design of SimpleITK[J]. Frontiers in neuroinformatics,2013,7:45.

[5] BYRD R H,LU P,NOCEDAL J,et al. A limited memory algorithm for bound constrained optimization[J]. SIAM journal on scientific computing,1995,16(5):1190-1208.

[6] GUTIERREZ-BECKER B,MATEUS D,PETER L,et al. Guiding multimodal registration with learned optimization updates [J]. Medical image analysis,2017,41:2-17.

[7] OTIRAS A, DAVATZIKOS C, PARAGIOS N. Deformable medical image registration: A survey[J]. IEEE transactions on medical imaging, 2013, 32(7): 1153-1190.

[8] BALAKRISHNAN G, ZHAO A, SABUNCU M R, et al. Voxelmorph: a learning framework for deformable medical image registration[J]. IEEE transactions on medical imaging, 2019, 38(8): 1788-1800.

[9] JADERBERG M, SIMONYAN K, ZISSERMAN A. Spatial transformer networks[J]. Advances in neural information processing systems, 2015, 28: 2017-2025.

[10] CHEN J, FREY E C, HE Y, et al. Transmorph: Transformer for unsupervised medical image registration[J]. Medical image analysis, 2022, 82: 102615.

[11] LIU Z, LIN Y, CAO Y, et al. Swin transformer: Hierarchical vision transformer using shifted windows[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal, QC, Canada, 2021: 10012-10022.

[12] CHEN Z, ZHENG Y, GEE J C. Transmatch: A transformer-based multilevel dual-stream feature matching network for unsupervised deformable image registration[J]. IEEE transactions on medical imaging, 2024, 43(1): 15-27.

[13] SUN J, SHEN Z, WANG Y, et al. LoFTR: Detector-free local feature matching with transformers[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, TN, USA: 2021: 8922-8931.

[14] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, USA: 2016: 770-778.

[15] DICE L R. Measures of the amount of ecologic association between species[J]. Ecology, 1945, 26(3): 297-302.

[16] HERNANDEZ-MATAS C, ZABULIS X, TRIANTAFYLLOU A, et al. FIRE: fundus image registration dataset[J]. Modeling and artificial intelligence in ophthalmology, 2017, 1(4): 16-28.

[17] 吴玲玉,蓝洋,夏海英. 基于卷积神经网络的眼底图像配准研究[J]. 广西师范大学学报(自然科学版),2021,39(5):122-133.

[18] RONNEBERGER O, FISCHER P, BROX T. U-net: Convolutional networks for biomedical image segmentation[C]//Medical Image Computing and Computer-Assisted Intervention-MICCAI 2015: 18th International Conference. Munich, Germany: Springer International Publishing, 2015: 234-241.

[19] KINGMA D P, BA J. Adam: A method for stochastic optimization[J]. arxiv preprint arxiv: 1412.6980, 2014.

[20] WANG Z, BOVIK A C, SHEIKH H R, et al. Image quality assessment: from error visibility to structural similarity[J]. IEEE transactions on image processing, 2004, 13(4): 600-612.

[责任编辑:陆炳新]