本科毕业设计（论文）报告

题 目：基于自监督学习的 2D/3D 脊柱图像配准研究

学　　号：09020119

姓　　名：陈闵恒

学　　院：计算机科学与工程学院

专　　业：计算机科学与技术

指导教师：孔佑勇

起止日期：2023/12-2024/05

东南大学毕业（设计）论文独创性声明

本人声明所呈交的毕业（设计）论文是我个人在导师指导下进行的研究工作及取得的研究成果。尽我所知，除了文中特别加以标注和致谢的地方外，论文中不包含其他人已经发表或撰写过的研究成果，也不包含为获得东南大学或其它教育机构的学位或证书而使用过的材料。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示了谢意。

论文作者签名：_____ 日期：_____年_____月_____日

东南大学毕业（设计）论文使用授权声明

东南大学有权保留本人所送交毕业（设计）论文的复印件和电子文档，可以采用影印、缩印或其他复制手段保存论文。本人电子文档的内容和纸质论文的内容相一致。除在保密期内的保密论文外，允许论文被查阅和借阅，可以公布（包括刊登）论文的全部或部分内容。论文的公布（包括刊登）授权东南大学教务处办理。

论文作者签名：_____ 导师签名：_____

日期：_____年_____月_____日 日期：_____年_____月_____日

摘　　要

基于图像的刚性 2D/3D 配准是荧光镜引导手术干预的关键技术。在骨科手术的过程中，通过将术前拍摄的 CT 图像与术中得到的 X 射线图像的坐标系进行对齐，从而得到病灶和手术器械当前在患者体内的精确的位置信息。近年来，伴随着机器学习技术的发展，许多基于学习的 2D/3D 配准方法被提出，相较于传统的基于优化的配准框架，基于学习的方法有着更少的运行时间，并且有着更好的全局搜索能力。然而，现有的基于学习的 2D/3D 配准研究大多都需要大量带有精确人工标注的数据来对神经网络进行训练，这需要大量的人力以及资源并且在这一任务中精确的标注是很难以实现的，从而限制了当前基于学习的方法的性能表现。本文提出了一种基于自监督学习的 2D/3D 配准框架，该方法通过采用自监督在线生成的合成数据进行训练从而免除了对大量带有标注真实数据的需求。整个框架由三个步骤组成，分别是基于深度回归的位姿初始化，基于深度度量的位姿搜索以及采用传统的基于优化的 2D/3D 方法的微调。本文分别在模拟的合成数据以及真实的成对的 X 光/CT 腰椎数据上，与现有的基于优化以及基于学习的基线方法进行了对比实验。实验结果表明本文提出的这一框架能够适用于单视角的 2D/3D 配准问题，并且验证了将基于优化的传统配准方法与基于学习的方法相结合对于配准结果的稳定性有着重要作用。

关键词：2D/3D 配准，图像引导介入，自监督学习

ABSTRACT

Image-based rigid 2D/3D registration is a critical technique for fluoroscopic guided surgical interventions. During orthopedic surgery, by aligning the coordinates of the pre-operative CT scans with the intra-operative X-ray images, the current precise position information of the lesion and surgical instruments in the patient's body can be obtained. In recent years, with the development of machine learning techniques, many learning-based 2D/3D registration methods have been proposed. Compared with conventional optimization-based registration frameworks, learning-based methods have less running time and better global search capabilities. However, most of the existing learning-based 2D/3D registration research requires a large amount of data with accurate manual annotation to train the neural network, which requires a lot of manpower and resources. And accurate annotation in this task is difficult to achieve, thus limiting the performance of current learning-based methods. This thesis proposes a 2D/3D registration framework based on self-supervised learning, which eliminates the need for a large amount of annotated real data by using synthetic data generated online under self-supervised training strategy. The entire framework consists of three steps, namely regression-based pose initialization, deep metric-based pose optimization, and refinement using traditional optimization-based 2D/3D methods. This thesis conducts comparative experiments with existing optimization-based and learning-based baseline methods on simulated data and real paired X-ray/CT lumbar spine data. Experimental results show that the framework proposed in this thesis can be applied to the single-view 2D/3D registration problem, and verify that the combination of traditional optimization-based registration methods and learning-based methods plays an important role in the stability of the registration results.

KEY WORDS: 2D/3D registration, Image-guided interventions, Self-supervised learning

目　　录

第一章 绪论

1.1 课题背景和意义

在基于 X 射线的图像引导的临床手术中，2D 的 X 射线透视图是提供图像引导的最佳方式。然而，由于所获得的二维图像的它本身的缺乏深度方向信息的性质，会导致存在一种固有的信息损失。而通过将术前的 3D 体积中的病灶以及人体器官叠加到 2D 图像上，可以在手术治疗期间提供必要的额外空间位置信息，从而能够有助于手术过程中的定位与导航。这一技术在当今新兴的采用手术机器人的微创骨科手术导航系统（如图 1-1 所示）中至关重要，在骨科手术中， 2D/3D 配准是一个旨在将术中 2D 图像（例如 X 射线图像）与相应的术前 CT 扫描对齐的过程。这一策略在当前的微创骨科手术导航系统中尤为重要，特别是在那些需要高精度导航的脊柱手术中，如经皮椎体成形术和椎弓根螺钉内固定术等。在这些手术中，2D/3D 配准技术发挥着核心作用，它通过将患者的实时解剖结构与预先获取的高分辨率 3D 图像进行准确匹配，使得外科医生能够在手术过程中实时观察到包括骨骼、神经、血管、植入物以及手术器械等在内的各种结构的位置和走向。这不仅极大地提升了手术的精确度，还有效保障了手术的安全性。

图 1-1 骨科手术机器人示意图

研究精度更高，速度接近实时的配准算法，能够有助于外科医生精确定位病灶位置，从而能够精确的切除病变部位，提高手术的成功率以及手术效率，减少医生的手术时间以减轻长时间手术带来的疲劳从而避免可能导致的术中的操作失误。相较于传统的介入式配

准方法，基于图像的配准技术为患者带来了近乎无额外损伤的优势。在手术过程中，外科医生仅需依赖全自动导航系统，该系统能自动完成配准并提供详尽的融合图像位置引导信息，无需医生进行额外的手动操作，具体流程示意图如图 1-2 所示。值得提到的是，现有文献中多数研究[1][2]均倾向于采用正位和侧位两个视角的 X 射线图像与术前 CT 进行配准。尽管多视角的 2D/3D 配准在技术上较为简便，且能较好地捕捉正位图像的深度信息，从而估算出人体与摄像头（成像光源点）之间的距离，但患者也因此需要接受更大剂量的辐射。本文的研究焦点在于单视角的 2D/3D 配准场景。这一场景相较于多视角更为复杂和具有挑战性，但它的优点包括所需要的成像资源需求更少且患者在成像过程中接受的电离辐射量也相应减少，这使得对单视角配准的研究具有重要的实际应用价值。

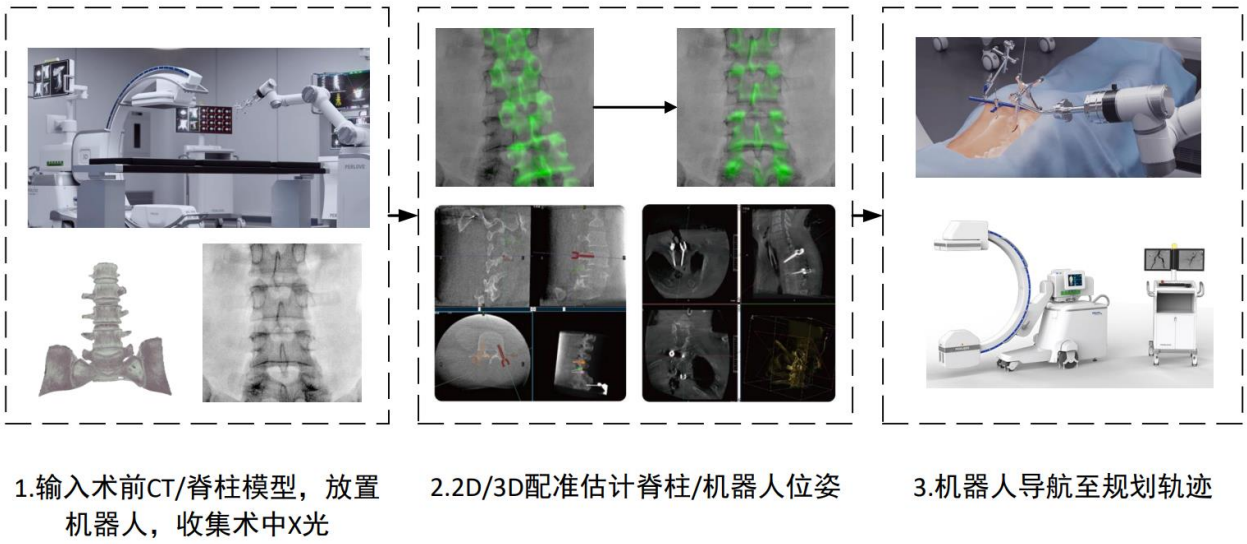

图 1-2 2D/3D 配准在骨科手术机器人导航中的应用示意图

近年来，伴随着机器学习技术的迅速发展，许多研究者尝试使用一些基于学习的方法来解决 2D/3D 配准问题。相较于传统的基于优化的配准方法，基于学习的方法有着极大的运行时间上的优势，在一些文献报告[3][4]中甚至达到了实时的运行速度，此外相较于容易陷入局部极小值的传统方法，基于学习的方法也有着更大的配准捕捉范围（捕捉范围指配准能够成功的最大位姿差异）。然而，现存的方法大多都是在大量的成对的 CT/X 光数据上进行训练的，这种数据要么是来自临床上采集到的来自同一病人的真实数据，要么则是通过重建 X 光图像获得其对应的 CT 扫描。模型的性能严重受到真实数据的数量以及质量的限制，且实际上关于本任务当前并不存在开源数据集。此外，采用真实数据进行训练还存

在获取到的 ground truth 的准确性难以保证的问题，通常研究者们会采用一些较为可靠完善的 2D/3D 配准处理流程[5]来估计真实数据的位姿，并进行人工验证和可视化检查，最后将这个结果作为真实数据的标签。也就是说，现有的真实数据的标签是某种或某几种基线方法的估计结果，从而可能导致数据本身的标签就是有偏差的，这也会影响最终训练出来模型的性能。

为了解决上述提到的问题，考虑到近年来模拟 X 射线图像生成技术的突飞猛进，本文提出使用自监督的训练策略，通过采用在线生成的模拟数据来对模型进行训练。此外，本文还提出了一种完备的 2D/3D 配准框架，以实现全自动的 2D/3D 脊柱图像配准。

1.2 研究现状

现有的 2D/3D 配准方法主要可以分为两个大类，传统的基于优化的方法和最近流行的基于学习的配准方法。其中基于优化的方法又可以被进一步细分为基于强度的方法和基于特征的方法两类，本节将依次对他们进行介绍。

1.2.1 基于优化的 2D/3D 配准方法

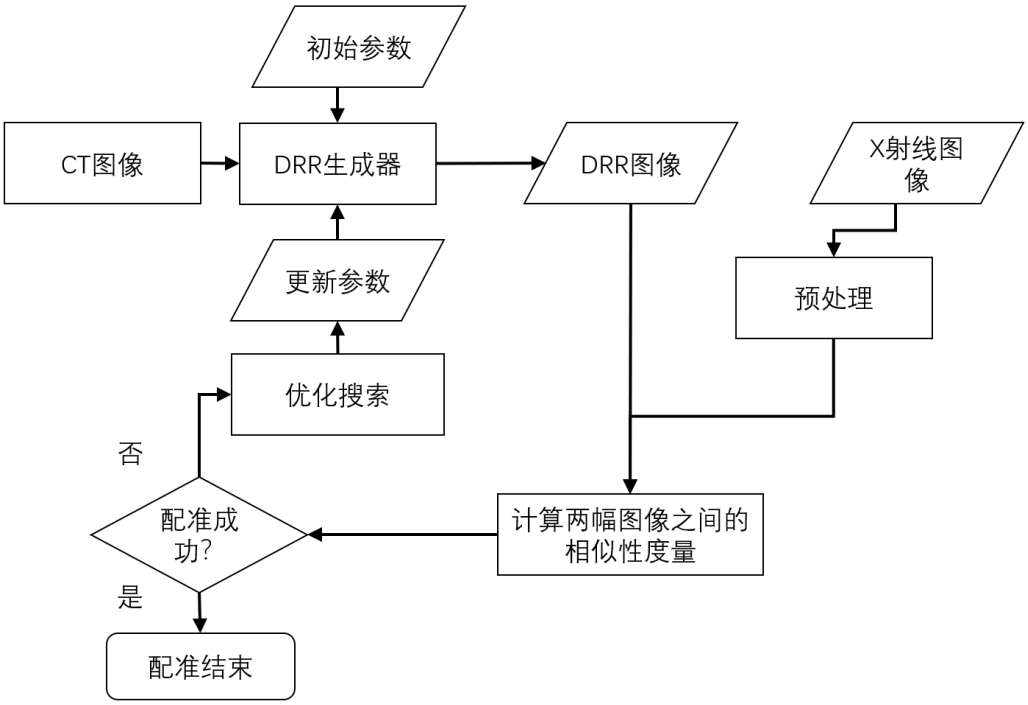

图 1-3 基于强度的 2D/3D 配准框架

基于强度的方法[6][7]的主体流程图如图 1-3 所示，其中 CT 首先经过数字重建放射影像（Digitally Reconstructured Radiograph，DRR）投影模块生成 2D 的 DRR 图像，再通过与经过预处理的 X 光透视图进行相似性测度的评估，来比较两张图像的相似性，如果两幅图像之间的相似性指标在达到某个预定标准则结束配准流程，此时 DRR 投影所使用的参数便是所需要求得的变换位姿；如果没有达到标准，则经过优化器进行寻优，找到一个新的最优解再进入 DRR 重新投影，再将结果与透视图做相似性测度比较。这里可以采用的优化方法和相似性度量有许多选择，有研究表明[7][8]不同的人体解剖部位所适应的相似性测度会有一定的差异，这主要是由于不同解剖部位的图像强度分布不同，所受到人体软组织的影响不同。如互信息（Mutual Information，MI）在软组织较多的部位上配准效果更佳，而归一化互相关（Normalized Cross Correlation，NCC）则更加适用于以骨组织为主的器官/部位。基于强度的方法的优点在于，它在初始姿态非常接近 ground truth 时，往往能取得一个非常高的精度，但它的缺陷则是较小的捕捉范围和较长的运行时间。

基于特征的方法[9][10]能够利用从图像中提取的几何特征（如角度、线条和分割）有效地计算相似性度量，因此比基于强度的方法具有更高的计算效率。现有的一些用于两组特征之间的匹配以及对应的位姿估计的算法，如 OpenCV 库中提供的 PnP，EPnP，PnP+RANSAC 等求解器能够很轻易的实现这种方法。然而，基于特征的方法存在一个潜在缺点，即它们严重依赖于几何特征的准确检测，这本身就是一项具有挑战性的任务。特征检测步骤中的误差不可避免地会传播到配准结果中，通常会损害基于特征的方法的准确性。相较于基于强度的方法，它有着更短的计算时间但也同样因此丢失了一些精度，但是这种方法的应用在不同类型数据上的鲁棒性往往难以保证。

1.2.2 基于学习的 2D/3D 配准方法

Miao 等人[3][4]率先将深度学习引入到了 2D/3D 配准领域，他们将整个配准看作是一个回归问题，即使用神经网络分别从移动的 DRR 与固定的 X 射线图像中提取特征并进行回归预测，这种方法的计算效率极高，能够达到低于 0.1s 的推理时间。在他们随后的工作[11]中将 2D/3D 配准的评估 DRR 与 X 射线图像之间姿态差异的过程看作是一个马尔可夫决策过程，从而采用一种利用基于全卷积网络（Fully Convolutional Networks）网络的多智能体强化学习算法进行配准。这两份工作被认为是基于学习的 2D/3D 配准方法的先驱。

在现有文献中，已经提出了两类主要的基于学习的 2D/3D 配准方法。第一类方法侧重于学习估计 2D 图像中的解剖标志或关键点与 3D 模型坐标之间的对应关系[12][13][14][15]。

在这种方法中，2D/3D 配准被建模为 Perspective-n-Points（PnP）问题，即在已知特征点在世界坐标系下的坐标以及特征点在成像平面的坐标，求解相机的位姿，而该问题已经有几个成熟的求解器可以直接使用[16][17]。基于解剖标志的方法由于其固有的稀疏性而具有较少的计算开销，但它的缺点是训练所需的特征和解剖地标是特定于每个对象的，并且必须通过专有的领域知识来选择。这表明获取大量精确标注的真实配对数据是跨多个领域的耗时且劳动密集型的过程，从而对该方法的临床适用性施加了重大限制。第二类方法通过离散位姿空间将问题转化为位姿回归问题[3][4][18]。 虽然这种方法可以实现实时配准并可用作姿态初始化策略，但由于透射 X 射线成像中直接映射的固有复杂性和不适定性，它在实现高精度姿态估计方面仍然受到限制。解决这个问题的常见方法是使用传统的基于强度的方法来细化姿势回归的预测姿势。此外，值得注意的是，许多基于回归的方法在有效集成输入的 2D 图像和 3D 体积的特征信息时面临挑战。为了解决这两种输入数据的维度差距，往往在特征融合时需要保证来自两个图像域中的特征数量相等，否则可能会导致模型不稳定并缩小配准捕获范围，甚至可能会使模型在推理时性能出现崩塌。现有的方法要么选择使用池化算子[19][20]将特征升维再直接融合，要么使用一些限制性机制从 2D 域重建空间信息，例如反投影[15][21]。这些方法的结果令人鼓舞，但过程缺乏可解释性。此外，单独使用 X 射线图像进行回归[22][23]对于非患者特定的任务是不可接受的，因为它忽略了来自 CT 体积的空间信息，导致训练域上的模型严重过度拟合。

最近，受传统基于优化的方法的启发，Gao 等人[24][25]提出了一种完全可微的框架，该框架通过使用神经网络来学习一种新的深度相似性度量用于 2D/3D 配准。该框架增加了配准的捕获范围，同时对局部特征的变化更具弹性。这种约束度量学习方法极大地减轻了大偏移和图像噪声引起的干扰，展现了极强的鲁棒性以及全局搜索能力。此外，一项初步研究[26]尝试使用相关驱动的方法在度量学习过程中分解局部和全局特征。

1.3 本文研究内容

基于上述内容，本文将致力于研究一种基于自监督学习的 2D/3D 配准方法，并提出一种可靠的全自动的 2D/3D 配准框架。如图 1-4 所示，框架流程分为三个阶段，首先，本文将对于位姿的初始化策略展开研究，提出采用一种直接的基于深度回归的方法来获取 CT 相对于 X 射线图像的位姿，接着，框架将得到的回归位姿作为基于神经网络的全局的位姿搜索优化过程的初值，多尺度的特征信息在许多相关文献中被验证对于配准的捕捉范围是

至关重要的，因此一种用于提取多尺度特征的复杂组合式编码器对于本课题而言是有迫切需求的。此外，虽然传统的基于优化的 2D/3D 配准是非常常见的基准方法，一个专门适用于本框架的基于优化的微调配准方法仍然需要被研究，对于微调步骤，本课题选择采用现有文献报告中表现最为优异的 2D/3D 配准优化算法之一的协方差自适应进化策略（Covariance Matrix Adaptation Evolution Strategy，CMA-ES）[27]作为优化器。除此之外，自监督学习最为重要的核心为模拟的 X 射线图像在线生成算法，经调研之后，本文虽然直接采用了现有的最先进的模拟成像方法：DeepDRR，但 DRR 的生成过程以及对应的在使用真实数据时针对 X 射线图像的预处理操作仍然需要研究与探索。本文中提出的方法的前两个模块，即基于深度回归的位姿初始化模块和基于深度度量的位姿搜索优化模块的工作已经以“Embedded Feature Similarity Optimization with Specific Parameter Initialization for 2D/3D Medical Image Registration”的题目发表在了 ICASSP2024（CCF-B）上，而关于基于深度度量的位姿搜索优化模块的可解释性和可控性的研究以及基于 CMA-ES 的微调的工作部分则发表在了 ISBI2024（医学影像领域顶级会议）会议上。

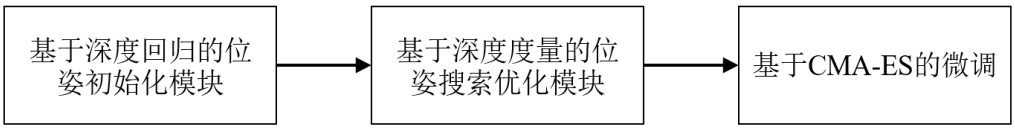

图 1-4 本文提出的三阶段的 2D/3D 配准框架流程示意图

综上所述，本文需要研究一个三阶段的 2D/3D 配准框架，其中包括基于深度回归的位姿初始化方法，基于神经网络和深度度量的位姿搜索优化以及基于传统方法即 CMA-ES 的微调。同时本文将介绍了相对应的对于真实 X 光数据的数据预处理过程和在线生成模拟 X 光数据的自监督训练策略的具体实现以及参数设置，最后还包括实验的设计和评价指标选取。

1.4 本文组织结构

本文将分为五个章节对本文的研究进行阐述，其中：

第一章，绪论部分，介绍本课题的研究背景、研究现状以及本课题研究内容。

第二章，相关技术与原理，介绍本文中利用到的相关技术与原理，包括 2D/3D 配准任务的定义，DRR 图像的生成和协方差矩阵自适应进化策略等。

第三章，框架模型设计与实现，介绍每个模块的的研究动机、模型框架、实现细节等。

第四章，实验设计与结构，介绍本文的实验环境、实验设计、实验结果以及参数实验。

第五章，总结与展望，介绍本文工作的大致内容，并且对目前研究的相关展望。

第二章 相关技术与原理

本章将围绕 2D/3D 配准的问题表述，数字重建放射影像的成像过程，常见的相似性测度，空间变换的约束方式以及协方差矩阵自适应进化策略展开，对本课题涉及的相关技术原理进行介绍。

2.1 2D/3D 配准问题表述

由于本课题主要关注的是应用于脊柱图像的 2D/3D 配准问题，而人体的椎节通常被看作是一个刚体，所以在本课题中只考虑基于刚体变换的配准。在三维空间中的刚体变换可以被一个六自由度的属于三维特殊欧式群的向量表示，即 $\theta \in \mathfrak{se}(3)$，在该向量中，有三个分量用于表示在 X、Y、Z 方向上的位移，即t_x、t_y和 t_z，三个分量被用于表示在这三个轴上的旋转，即r_x、r_y和 r_z。

刚体的 2D/3D 配准问题（如图 2-1 所示）可以被看作是一个优化下列式子的过程：

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}}\, S(I, P(\theta; V)) \tag{2.1}$$

其中I是固定的 2D 术中 X 射线图像，S是一种评估两幅图像/向量间相似度的度量方式，本课题中既采用了基于统计的相似性函数也采用了基于神经网络的一种映射来评估图像的相似性。$P(\theta; V)$则是表示三维体积V在经过了六自由度的刚体变换θ之后，再通过 DRR 生成器$P(;)$进行投影操作得到了一张 2D 的 DRR 图像。

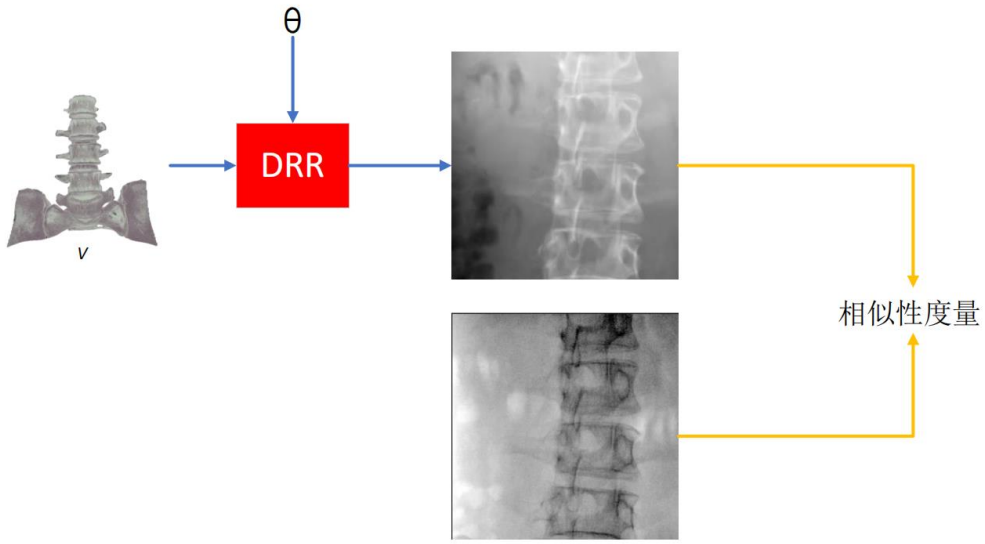

图 2-1 2D/3D 配准优化目标示意图

2.2 数字重建放射影像的成像过程

数字重建放射影像（Digitally Reconstructured Radiograph，DRR）是使用各种射线追踪技术从 3D 的计算机断层扫描（CT）体积生成的二维 X 射线图像。如何从一张三维的 CT 扫描中模拟生成出一张 X 射线图像是一个相当困难的问题，虽然 CT 本身的成像原理与 X 射线图像一致，都是由 X 线穿过人体衰减得到的信号生成，所以成像理论上是可行的，但因为对于这种算法的需求往往都是在手术中的一些临床上的应用，因而对于成像的需求不仅是要求要有较高的真实性，同时也需要较快的投影速度。Siddon 方法[28]是最常用的用于 DRR 合成中的光线追踪方法，近年来也有许多扩展研究致力于在 GPU/CPU 上提高这种算法的渲染速度，如 Siddon-Jacobs'方法[29][30]，它通常在 CUDA 和 C++中实现，以创建多线程 GPU 加速的 DRR 生成器，通过分配每个线程来跟踪与探测器平面相交的独立射线来利用数据并行性。

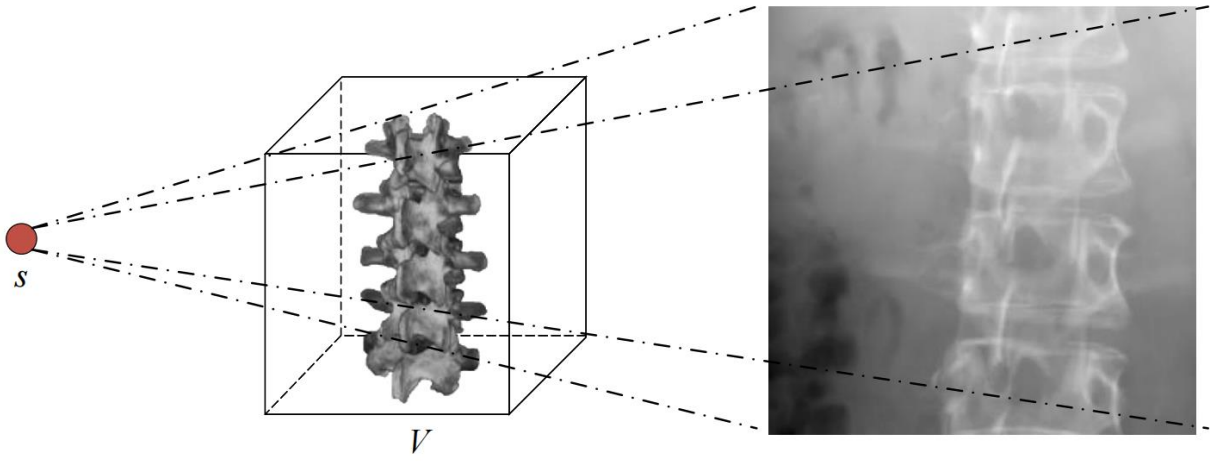

图 2-2 DRR 生成示意图

假设$s \in \mathbb{R}^3$为 X 射线的光线源，$p \in \mathbb{R}^3$为探测板平面上的目标像素。然后$R(\alpha) = s + \alpha(p - s)$是一条来自$s\ (\alpha = 0)$的射线，穿过了被成像的物体，并在$p\ (\alpha = 1)$处射在探测板平面上。X 射线在达到像素$p$时所经历的总能量衰减由以下直线积分给出：

$$E(R) = \| p - s \|_2 \int_0^1 V\,(s + \alpha(p - s))\, d\alpha \tag{2.2}$$

其中V: $\mathbb{R}^3 \rightarrow \mathbb{R}$为成像体积。$\| p - s \|_2$则是用于赋予无单位长度的$d\alpha$的单位物理长度。因为这里的$V$采用的是 3D 的 CT 体积来近似，它并不是真正的连续的含有不同密度的物质成分的人体而是一个离散的三维体积，从而公式 2.2 需要被修改为离散化的表示形式：

$$E(R) = \| p - s \|_2 \sum_{m=1}^{M-1} (\alpha_{m+1} - \alpha_m) V\left(s + \frac{\alpha_{m+1} + \alpha_m}{2} (p - s) \right) \tag{2.3}$$

其中，α_m参数化了射线R与一个只包含 CT 体积的正交平面相交的位置，而 M 是这些相交点的数量。通常来说，M 的值越大，生成的图像就显得越真实，但是值得注意的是，上述提到的这种 DRR 成像方法，它将 X 线的衰减模拟为一种均匀的衰减，即它假设的是穿过的物体的密度是均匀的，而显然在真实的环境中，软组织、骨骼以及人体内的空气和水对于 X 射线的吸收率是完全不一样的，且这里也并没有考虑在真实的 X 线成像系统中可能出现的光子的反射与漫射，以及探测板可能收到光子噪声的影响。

为了考虑上述提到的这几种因素的影响，生成出更加逼真的合成 X 光图像，约翰霍普金斯大学的研究者们提出了 DeepDRR[31][32]，当前最先进的基于蒙特卡洛模拟的用于从 CT 扫描快速得到真实模拟透视图像和数字放射成像的框架。DeepDRR 的主要流程包括如下几个步骤：1）首先对于 CT 扫描进行一个体素级别的材料分解（Material Decomposition），将 CT 中的骨骼、软组织、空气用神经网络分解出来。2）根据三幅材料的衰减图，采用蒙特卡洛模拟法，模拟光子穿过人体的投影过程。3）对投影图像进行基于学习的散射估计。4）向图像添加泊松噪声来模拟 X 光成像系统中由光子统计量引起的不相关量子噪声组成的组合（这种量子噪声由于像素串扰和相关读出噪声成为相关的），即首先计算每个像素的平均光子能量从而生成对应的泊松噪声模型下的量子噪声，接着通过将噪声信号与一种模糊核进行卷积，将相邻像素的量子噪声关联起来从而模拟真实的探测板上的像素串扰，最后再对图像沿行相关的加性高斯噪声以模拟图像的电子读出噪声（Electronic Readout Noise）。本课题采用 DeepDRR 来用于生成自监督训练所需要的模拟 X 光数据。

近年来，伴随着机器学习在本领域应用的需求，一些研究者们提出了一些全可微的 DRR 生成器，如 ProST[25]和实现向量化 Siddon 方法的 DiffDRR[33]，本研究的神经网络内部以及计算无监督的损失函数所用到的 DRR 生成器均采用的是最近被提出的 Projective Spatial Transformer（ProST）。

2.3 常见的相似性度量

常见的用于刚性配准的相似性度量主要都是基于统计的方法，如互信息（MI）或者归一化互相关（NCC）。这些方法旨在通过最小化图像对之间联合直方图的熵来估算一种解决方案，从而实现精确的刚性配准。接下来本小节将依次介绍这些相似性度量以及它们的变种。

$$NCC(I_1, I_2) = \sum_{i=0}^{m} \sum_{j=0}^{n} \frac{(I_1(i,j)-\bar{I_1})(I_2(i,j)-\bar{I_2})}{\sigma_{I_1}\sigma_{I_2}} \quad (2.4)$$

公式 2.4 描述了两幅大小皆为$m \times n$的 2D 图像之间计算互相关的过程，其中$\bar{I_1}$和$\bar{I_2}$分别为两幅图像的灰度均值，σ_{I_1}和σ_{I_2}则为它们各自的标准差。在互相关中，给定像素的贡献强烈依赖于像素的强度，因此，强度上的一些较大的大差异（如可能由介入器械遮挡引起）会对相似度度量产生重大影响。而相比而言，互信息并不假设两幅图像的像素值之间存在线性关系，而是假设两幅图像在配准成功时最可能值的共存程度最大。因而这种相似性度量非常适用于 3D-3D 的多模态配准任务，如 MR/CT 或 PET/CT 配准。MI 的计算公式如下所示：

$$MI(I_1, I_2) = \sum_{x,y} p(x,y) log \frac{p(x,y)}{p(x)p(y)} \quad (2.5)$$

其中$p(x)$和$p(y)$分别表示两幅图像各自的概率分布，$p(x,y)$则表示它们的联合概率分布。

为了去除软组织对于图像强度的影响，基于 Sobel 算子的梯度相关（Gradient Correlation，GC）被提出了。梯度相关通过计算微分来改变 I_1 和 I_2的强度分布。首先利用水平和垂直 Sobel 模板生成梯度图像 dI_1/di 和 dI_1/dj，表示沿图像两个正交轴的透视强度的导数。随后，计算dI_1/di和dI_2/di之间以及dI_1/dj和dI_2/dj之间的归一化互相关。该度量的最终值是这些归一化互相关的平均值。具体公式如下所示：

$$GC(I_1, I_2) = \frac{NCC(dI_1/di, dI_2/di) + NCC(dI_1/dj, dI_2/dj)}{2} \quad (2.6)$$

然而虽然基于梯度的度量可能对软组织的形变不再敏感，但 GC 的 Sobel 算子的影响会导致它对细线形的结构敏感。为了解决这个问题，梯度差分（Gradient Difference，GD）[8]被提出了，它在 GC 的基础上进一步改进，采用梯度图像的差图的熵的形式来度量图像的相似性。具体公式如下：

$$GD(I_1, I_2) = \sum_{i,j \in I} \frac{A_x}{A_x + (dI_1/di(i,j) - s \cdot dI_2/di(i,j))^2} + \sum_{i,j \in I} \frac{A_y}{A_y + (dI_1/dj(i,j) - s \cdot dI_2/dj(i,j))^2} \quad (2.7)$$

其中A_x和A_y为固定的常量，通常等于差图的标准差，而s则是一个常数用于统一两幅图像的灰度分布的尺度。

基于块的 NCC 方法近年来也不断被人们使用，有的文章[34]将它称之为局部互相关（local NCC）,这种度量的计算方式是先将图像切为K个可重叠/不重叠的以点$(p_x, p_y) \in \Omega_K$为中心，r为半径的小块。小块内的 NCC 计算公式如下：

$$LNCC(I_1, I_2, p_x, p_y, r) = \sum_{i=p_x-r}^{p_x+r} \sum_{j=p_y-r}^{p_y+r} \frac{(I_1(i,j)-\bar{I_1})(I_2(i,j)-\bar{I_2})}{\sigma_{I_1}\sigma_{I_2}(2r+1)^2} \quad (2.8)$$

其中σ_{I_1}和σ_{I_2}是局部的块内的像素的标准差而不是如公式2.4中的那样计算整幅图像的标准差。

最近一些研究[22][33][35]表明将全局的NCC与局部的LNCC相结合，能够得到与GC以及Patch-GC（patch-based gradient correlation）[6]相比相对更加平滑的相似性函数，这更有利于在人体位姿有较大偏移的情况下捕捉到在解空间内的最优值，相较而言GC以及Patch-GC它们的函数曲线虽然在最优值附近更加尖锐，但全局上它们相对更加不平滑，有着更多的局部极小值，从而使得当患者的位姿偏移较大时配准失败的可能性增加。这种提出的度量被称作为多尺度的归一化互相关（multiscale Normalized Cross Correlation，mNCC）。

$$mNCC(I_1, I_2) = NCC(I_1, I_2) + \lambda \sum_{(p_x,p_y)\in\Omega_K} LNCC(I_1, I_2, p_x, p_y, r) \quad (2.9)$$

本研究所采用的局部的基于块的NCC仅考虑了块之间不重叠的情况，重叠情况下的LNCC[6] [22]相对而言所需要的计算资源更多，虽然可以采用稀疏化渲染的方式进行加速但会导致所估计的相似性函数是有偏的，即对配准结果的影响难以估计以及其稳定性难以保证，所以在本课题中并没有采用。

文献[8]中的实验结果表明，并不存在一个相似性度量能够在所有人体部位的数据上都取得最好的配准结果，所以通常研究者们都需要针对于不同的人体部位选择一个最为合适的相似性度量。如相较于NCC和MI，GC和GD对于软组织存在所造成的干扰的影响较小，当透视图像中存在金属支架时，基于熵的直方图统计方法，即GD受到的影响则较小，当软组织和金属支架的影响在透视图像中同时存在时，使用最常见的基于统计的相似性度量NCC以及MI都会直接出现较大的配准失败，而GD和GC等变体的表现则相对更为稳定。

2.4 空间变换

2D/3D配准的目标就是为了找到浮动图像相对于固定（参考）图像的位姿的变换参数，在三维空间中的空间变换的约束形式有许多种，如刚体变换，仿射变换，投影变换以及一些高维的基于生物物理学/生物力学或粘性流体模型的弹性形变（deformation）。在本文中，因为配准的目标为人体的脊柱，这是一个由多节紧密连接的骨组织组成的刚体结构，所以本研究将仅关注于六自由度（6 DoF）的刚体变换。

对于刚体变换的参数$\theta \in \mathfrak{se}(3)$来说，它可以被拆分为旋转和平移两个部分，即旋转分量$r = \{r_x,\ r_y,\ r_z\}$和平移分量$t = \{t_x,\ t_y,\ t_z\}$。三维空间中的旋转通常由一个3×3的矩阵R表示，其中这 9 个元素受 6 个范数和正交约束（R是正交矩阵且$\det(R) = 1$）。而前面所提到的三个自由度的旋转向量r，它是一个嵌入在$\mathbb{R}^3$空间中的三维对象（因此有 3 个自由度）集合名为特殊正交群（Special Orthogonal Group）$\mathrm{SO}(3)$。根据欧式几何学，每一个旋转矩阵R都可以由这样的一个三维的向量来进行表示。它们之间的转换过程经过罗德里格斯（Rodrigues'）旋转公式简化后所示如下：

$$\hat{r} = \frac{r}{\|r\|_2} \tag{2.10}$$

$$[\hat{r}] = \begin{bmatrix} 0 & -\hat{r}_z & \hat{r}_y \\ \hat{r}_z & 0 & -\hat{r}_x \\ -\hat{r}_y & \hat{r}_x & 0 \end{bmatrix} \tag{2.11}$$

$$R = I_3 + \sin(\|r\|_2)\,[\hat{r}] + (1 - \cos(\|r\|_2))[\hat{r}]^2 \tag{2.12}$$

因此，要在三维空间中找到任意的一个旋转朝向，只需要找到该方向对应的旋转向量r就足够了。从而整个的三维的$\mathrm{SE}(3)$空间中的物体V的位姿变换可以被表示为：

$$V' = R \cdot V + \begin{bmatrix} t_x \\ t_y \\ t_z \end{bmatrix} \tag{2.13}$$

在本文中$\theta \in \mathfrak{se}(3) \longrightarrow \begin{bmatrix} R & t \\ 0 & 1 \end{bmatrix} \in \mathrm{SE}(3)$的过程被定义为$\begin{bmatrix} R & t \\ 0 & 1 \end{bmatrix} = exp(\theta)$，即用$exp(\theta)$表示六自由度的位姿向量变换为变换矩阵的过程。

2.5 协方差矩阵自适应进化策略

进化策略是一种基于自然选择思想的生物启发算法，而协方差矩阵自适应进化策略[27]（CMA-ES）则是当前最为先进的它的改进。CMA-ES 算法的优化过程可以被视作是一个不断采样的过程。CMA-ES 使用多元正态分布生成候选解来最小化目标函数 $f:\ \mathbb{R}^d \longrightarrow \mathbb{R}$，这个分布$N$由三个元素参数化：均值向量 $m \in \mathbb{R}^d$、步长 $\sigma \in \mathbb{R}$ 和协方差矩阵 $\pi \in \mathbb{R}^{d \times d}$。

在迭代$t+1$中，首先根据当前分布$N(m^{(t)}, \sigma^{(t)2}\pi^{(t)})$独立采样$\omega$次得到$\omega$个候选解$x_i^{(t)}$及其关于目标函数的对应值$y_i^{(t)} = f(x_i^{(t)})$。接下来，根据采样得到的当前代的候选解计算分布的演化路径，最后更新分布参数，得到演化分布$N(m^{(t+1)}, \sigma^{(t+1)2}\pi^{(t+1)})$ 。这当中复杂的参数更新过程可以被简化为如下的式子：

$$m^{(t+1)} = m^{(t)} + \Delta_m^{(t)} \tag{2.14}$$

$$\Sigma^{(t+1)} = \Sigma^{(t)} + \Delta_\Sigma^{(t)} \tag{2.15}$$

其中$\Sigma^{(t)}$是$\sigma^{(t)2}\pi^{(t)}$的向量化表示。CMA-ES 已经被许多研究应用[36]在 2D/3D 配准这个任务中并且都表现出了不错的表现。进化策略优化目标函数f就是图像的相似性测度，即计算$S(I, P(\theta; V))$（公式 1.1），d设置为刚性形变变换参数的维度数为 6，初始的向量$m^{(0)}$则为配准时设定的初始位姿，σ和π则可以为手动设置的变量。在迭代进化了T代之后，程序终止，再将得到的$m^{(T)}$作为搜索得到的最终结果输出，它也是配准估计得到的变换参数。CMA-ES 的优点包括较快的收敛速度以及较强的全局搜索能力，它能够通过自适应的调整步长和协方差矩阵的方式，来保证搜索到全局最优解。但它的缺陷也相对比较明显，首先是几个超参数代数T以及每代的采样数ω在不同的实验条件下需要反复调整来找到最优的参数设置。此外较长的搜索时间也使得这种方法在手术中的临床应用受到了较大的限制，在许多初期研究报告中[7][36]，采用 CMA-ES 的配准流程平均时间超过了两分钟，虽然近年来伴随着 GPU 加速技术的发展以及 CUDA 算力的提升，整体的配准时间能够达到低于一分钟[37]，甚至在一些最新的报告中它可以在 20s 内完成配准[24]，但相较于许多声称能够达到实时的基于学习的方法[38]而言它的计算效率仍然有待提升。

2.6 本章小结

在本章中，首先介绍了 2D/3D 配准任务的定义以及表述，接下来介绍了 DRR 图像的成像原理以及一些现有的相关工作，之后引入了许多在传统的基于优化的方法中被大量采用的一些相似性度量和本课题所采用的空间变换方式，最后本章介绍了 CMA-ES 算法的具体细节以及它在 2D/3D 配准领域中的应用。本章通过介绍以上的相关技术为后续的研究提供理论基础。

第三章 基于自监督学习的 2D/3D 配准框架

本章主要研究一种基于自监督学习的 2D/3D 配准框架。3D 浮动图像是术前待配准的 CT 图像，2D 参考图像是通过 DeepDRR[31][32]技术模拟成像得到的 X 光图像或真实的术中获得的 X 光图像。本章首先介绍本文设计并实现的一套基于自监督学习的 2D/3D 配准框架的总体结构和流程，然后分别介绍该框架的三个子模块：基于深度回归的位姿初始化方法，基于神经网络和深度度量的位姿搜索优化以及基于 CMA-ES 的微调。最后也介绍了在自监督训练时使用的域随机化数据增强方法的具体操作。

3.1 基于自监督学习的 2D/3D 配准总体流程

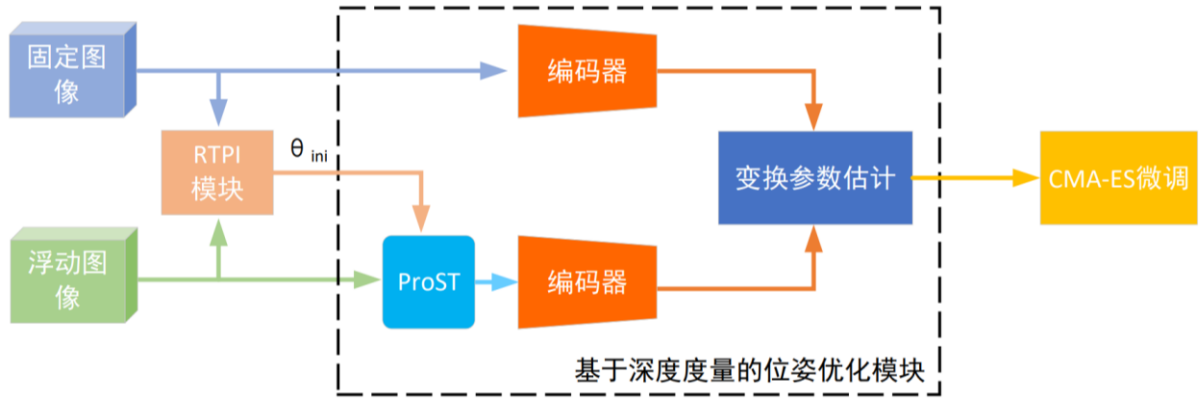

图 3-1 基于自监督学习的 2D/3D 配准总体框架

本文所提出的基于自监督学习的 2D/3D 配准框架如图 3-1 所示。框架的输入为 3D 的术前拍摄的 CT 扫描（即浮动图像），2D 的术中得到的 X 光图像（即固定图像），首先将两幅图像输入到本文提出的基于深度回归的位姿初始化模块，刚体变换的位姿初始化网络（Rigid Transformation Parameter Initialization，RTPI），该网络的输出为一个初始的位姿θ_{ini}，接着得到的初始位姿和浮动还有固定图像将被一起输入至迭代式的基于深度度量的位姿搜索优化模块（Iterative Fine-registration Module，FineReg），如图 3-1 中所示，该模块主要由一个 DRR 的生成器（ProST[24][25]）和两个权重独立的编码器组成，在每一次迭代中，ProST 模块根据当前的位姿θ_i投影得到一张 DRR 图像$I_m^{(i)}$，再根据深度度量提供的相似性采用梯度下降法优化当前的位姿θ_i，更新得到新的位姿θ_{i+1}。因为经过实验发现，虽然这种基于深度学习的度量方法的相似性函数曲线比起传统的相似性度量更为平滑，且在更大的范围上是凸的，有着更强的全局上的信息捕捉能力从而增大了配准的捕捉范围，但是它在接近全

局最优解即 ground truth 附近时会过早的收敛，从而导致估计的位姿在精度上表现不佳，虽然本研究引入了一种能够提取多尺度特征的骨干网络作为编码器来缓解这一缺陷，该问题仍然没有很好的得到解决。因此在得到了基于深度度量的优化得到的位姿结果之后，本研究还使用了一种基于 CMA-ES[27]的传统的 2D/3D 配准方法来微调前者的结果以提升预测结果的精度并保证整体框架的鲁棒性。需要值得注意的是，本文所说的自监督训练策略是指，本文所有神经网络都是在采样生成的位姿对投影得到的模拟合成数据上进行的训练。

3.2 基于深度回归的位姿初始化模块

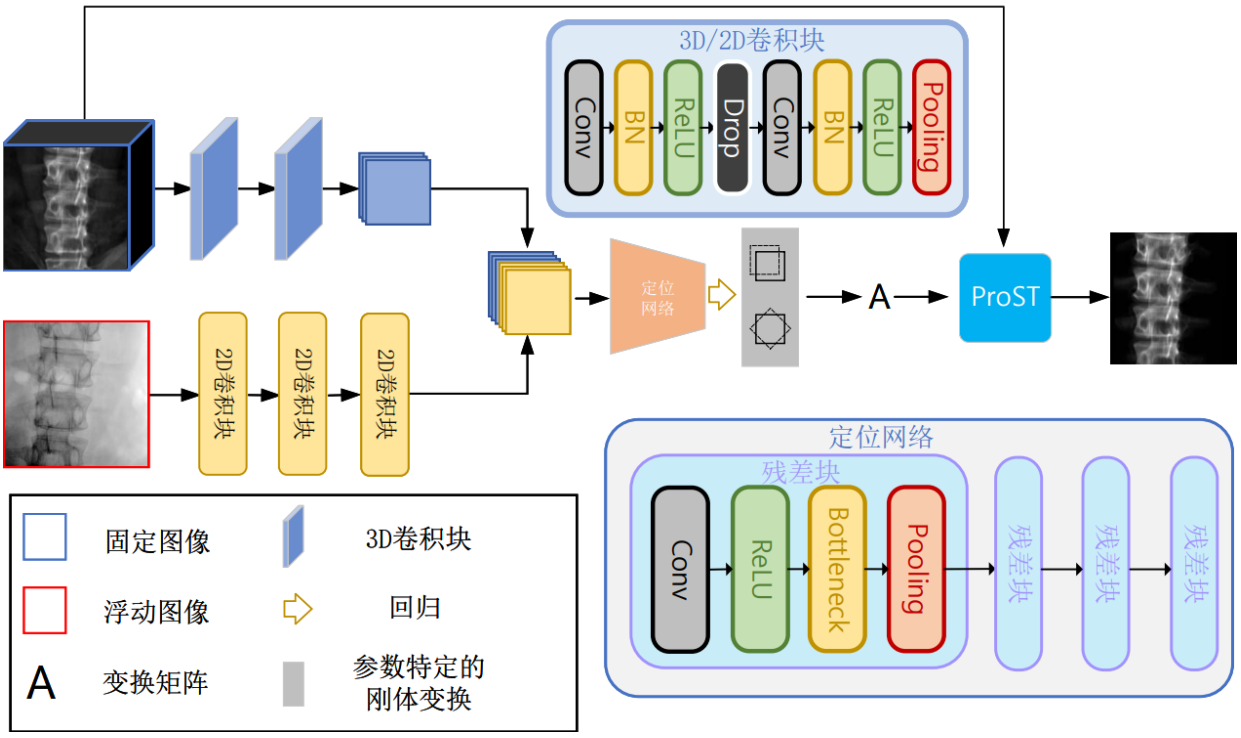

图 3-2 刚体变换的位姿初始化网络（RTPI）示意图

本节将介绍所提出的基于深度回归的位姿初始化模块，该模块的结构示意图如图 3-2 所示。网络的输入为一个 3D 的体积V和一个 2D 的固定图像I_f。两个输入数据之间的维度差异对于配准的表现有着非常重大的影响，因而融合来自这两个不同维度的图像域的信息的方法至关重要。针对提取来自两个不同的图像域的特征并将它们进行融合，通常有两种办法。第一种方法便是类似 Voxelmorph[39]的特征早融合，即在将两幅图像输入网络之前直接在通道维度上拼接在一起，这种方法在实现上非常简单且在 3D/3D 或者 2D/2D 配准任务中被验证是有效的，但由于在 2D/3D 配准这个任务中，数据的维度和信息量是不对称的，所

以早融合会导致网络忽视来自 2D 图像域的信息内容，从而导致实际回归的结果与 ground truth 的差值较大，因为实际上 CT 扫描的灰度分布上的变化大多是病人本身身体解剖结构上的差异，这种变化或者差异对预测的位姿的影响在设想中不应该比 2D 图像域的特征要大，所以早融合的方法是不适用于此任务的。第二种融合方法便是晚融合，即先用两个分支的神经网络分别去提取两幅图像的特征，再将它们拼接在一起进行回归预测。在一些先前的工作中[19]，在处理 2D 和 3D 的跨维度数据时，研究者们采用的方法是在提取 2D 图像特征的分支的最开始使用池化操作，将特征的通道维度变为与 3D 图像的深度（D）的维度数一致，这样两个分支后续都只需要采用三维卷积来分别提取特征再融合起来，这种方法被验证在超声图像的切片到体积（slice-to-volume）的配准中是有效的，但它的缺点便是因为网络整体都是采用的三位卷积实现，所以在推理速度和参数量上相对较大。为了避免这一问题，与该方法不同的是，在所提出的网络中三维卷积被采用来提取 3D 体积的特征，二维卷积被用于提取 2D 图像的特征，接着在特征融合之前将三维域的特征图展开，特征向量的尺寸从(C,H,W,D)的变为$(C,4H,4W)$。提取 3D 特征的网络分支共有两个卷积块，提取 2D 特征的网络分支则有三个卷积块，每个卷积块的结构都是完全一致的，即由两个 2D/3D 卷积层-BN 层-ReLU 层的序列中间间隔着一个 dropout 层组成，并在最后有一个平均池化操作。再将来自两个分支的等同数量的特征进行拼接之后，再经过一个由三个残差瓶颈[40]和最大池化层组成的定位网络对特征降维，再经过三个输入输出维度分别为$(1024,512)$，$(512,128)$和$(128,6)$的全连接层，最后经过一个取值在[0,1]之间的 ReLU 函数。这里将 2D 分支网络和 3D 分支网络分别用$\Phi_{2D}(\cdot)$和$\Phi_{3D}(\cdot)$表示，用于特征降维的定位网络为$L(\cdot)$，则整个回归获得回归结果x的过程可以表示为：

$$x = ReLU_{[0,1]}(L(concat(\Phi_{2D}(I_f), \Phi_{3D}(V)))) \tag{3.1}$$

因为之前有文献表明[41]，相较于直接回归4×4的变换矩阵$\begin{bmatrix} R & t \\ 0 & 1 \end{bmatrix} \in \mathrm{SE}(3)$，回归其对应的$6 \times 1$的向量表示$\theta$有助于减少变换的二义性，并且也会降低深度回归的拟合难度。具体的回归方法采用的是一种参数特定（parameter-specific）[42]的形式，即对于每一个要回归的位姿参数θ_i，都被手动的设置了它在解空间中的上下界限θ_i^{min}和θ_i^{max}。因为x_i的值域为[0,1]，所以计算回归每一个位姿分量θ_i可以写为下列式子：

$$\theta_i = \theta_i^{min} + x_i\left|\theta_i^{max} - \theta_i^{min}\right| \tag{3.2}$$

这种方法限制了深度回归的解空间的范围，能够很好的加速训练的收敛过程，同时避免了预测的初始位姿与 ground truth 有较大误差从而影响后续的流程以及配准的最终结果。后续为了采用 ProST 进行投影获得 DRR 图像I_m，所以θ会被根据公式 12 转化为对应的4×4的变换矩阵 **A**。

在许多无监督图像配准方法中，传统的基于优化的配准（例如 2.3 小节中提到的 NCC、MI、GC）的相似性度量被成功地用作梯度下降优化的损失函数。然而，相似性度量的缺点是很容易陷入局部极小值。因此可以计算监督均方误差（mean square error，MSE）损失，它直接使用真实标签$\hat{\theta}$的信息，能够更加有效的引导网络对深度回归预测进行拟合。

$$L_{mse}(\hat{\theta},\theta) = \frac{1}{N}\sum_{i=1}^{N}\left\|\hat{\theta}_i - \theta_i\right\|_2 \tag{3.3}$$

通过上面提到的均方误差和相似度损失，刚性变换参数初始化网络的参数根据以下损失函数的引导更新：

$$L_{RTPI} = \alpha L_{sim}(I_m, I_f) + \beta L_{mse}(\hat{\theta},\theta) + \lambda\mathcal{R}(\theta) \tag{3.4}$$

其中L_{sim}是在本文 2.3 小节中提到的无监督的相似性损失梯度差分（GD），$\mathcal{R}(\theta)$是一个 L2 正则化项，其正则化参数为 λ。另外，α、β是超参数。

3.3 基于深度度量的位姿搜索优化模块

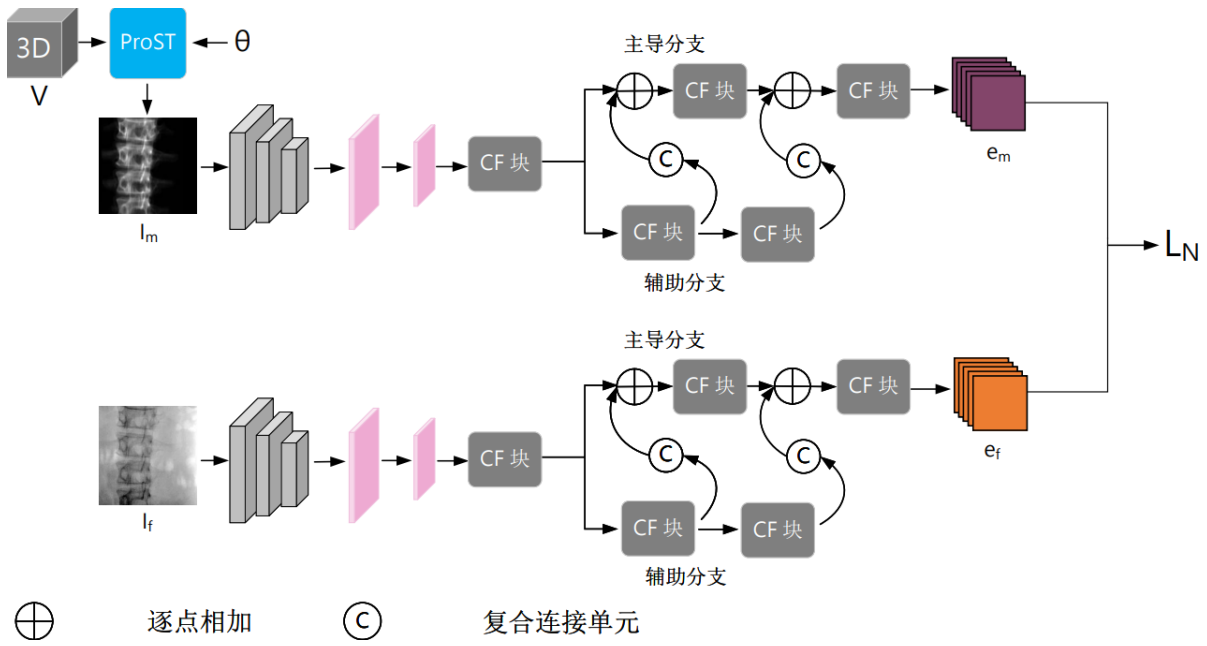

图 3-3 基于深度度量的位姿搜索优化模块示意图

本节将介绍所提出的基于深度度量的位姿搜索优化模块，该模块的结构示意图如图 3-3 所示。网络的输入为一个 3D 的体积V和一个 2D 的固定图像I_f以及由 RTPI 网络得到的预测的初始位姿 θ_{ini}，这里可以将它写作是$\theta^{(0)}$。网络的正向传播大致为首先 ProST[25]相对于 $\theta^{(t)}$ 进行投影操作。投影的运动图像I_m和输入的固定图像I_f各自经过两个结构相同但不共享权值的编码器，输出嵌入特征e_f和 e_m。

因为该模块是为在预测的初始位姿的基础上进行精细配准而设计的，因此设计一种多尺度特征融合的网络架构至关重要，该架构可以提取图像的细粒度纹理特征，同时关注高级特征，从而扩大捕获范围。提取多尺度的特征对于 2D/3D 配准的好处在无论是在传统的基于强度的方法[6]还是在基于学习的方法中都被论证过[24]，在此本文就不再重复论述了。本研究吸收了级联融合网络（Cascade fusion network，CFnet）[42]和复合骨干网络（Composite Backbone Network，CBnet）[43]这两种网络架构的优点，提出了一种用于多尺度特征提取和融合的复合编码器（Composite Encoder，CE），它可以提取语义和全局信息转化为嵌入式功能。编码器由两个分支组成，提取低级特征的辅助分支将通过复合连接单元和连接操作融合到主导分支。因此，在主导分支上，可以获得能够更好地表示图片之间相关性的高层和低层特征。在将编码器分成两个分支之前，首先使用三层卷积网络从图像中提取浅层特征，在每层卷积之后，有一个激活函数和一个归一化层。然后接下来使用两个卷积层对特征进行下采样，然后再经过一个级联融合块，之后分别送入两个分支中。每个复合连接单元均由 1×1 卷积层和用于减少通道的批量归一化层以及上采样操作模块组成。在每个分支上都有两个级联融合块，与原文中[42]的实现有所区别的地方在于在此处采用了 spilt-attention bottleneck[44]作为级联融合块的瓶颈。

该模块的训练策略是基于测地线距离，一种被许多研究表明在黎曼几何中的特殊欧式群（SE）以及特殊正交群（SO）上是凸的距离度量。Mahendran 等人[45]率先将它应用到姿态估计任务上，但他们只用它来计算旋转分量间的误差，即在$\mathrm{SO}(3)$上的测地线距离。随后一些研究[46][47]将它引入到医学影像领域。给定两个姿态θ_1和θ_2，根据公式 12 可以得到它们对应的旋转矩阵R_1和R_2，它们在$\mathrm{SO}(3)$的流形上的测地线距离可以被表示为：

$$d(R_1, R_2): \mathrm{SO}(3) \times \mathrm{SO}(3) \longrightarrow \mathbb{R}^+ = \|log(R_1^T R_2)\|_F \tag{3.5}$$

其中$\|\cdot\|_F$是弗罗贝尼乌斯范数（Frobenius Norm），$log\ (\cdot)$是矩阵对数计算操作，对矩阵R的对数化操作可以写成如下形式：

$$log(R) = \begin{cases} 0 & if\ \|r\|_2 = 0 \\ \frac{\|r\|_2}{2\sin(\|r\|_2)}(R - R^T) & if\ 0 < \|r\|_2 < \pi \end{cases} \tag{3.6}$$

又因为已知实际上：

$$tr(R) = 1 + 2\cos(\|r\|_2) \tag{3.7}$$

则公式 20 可以被改写为：

$$d(R_1, R_2) = \cos^{-1}\left[\frac{tr(R_1^T R_2) - 1}{2}\right] \tag{3.8}$$

将上述的式子扩展至$\mathrm{SE}(3)$中则可以得到：

$$L_{geo}(\theta_1, \theta_2) = d(T_1, T_2) = \|log(T_1^T T_2)\|_F \tag{3.9}$$

其中$T_1 = exp(\theta_1)$，$T_2 = exp(\theta_2)$。（见 2.4 小节）

本文采用了在[24][25]提出的两次反向传播（Double-backward）的训练机制来训练网络，该机制与一些其它的深度度量近似方法[48]不同，没有选择直接让网络去近似在$\mathrm{SE}(3)$中的测地线距离，而是让网络的梯度去近似测地线距离的梯度，这种二阶的近似方法有效的摆脱了距离的尺度对于网络近似效果的影响。具体的训练过程，在第t次的迭代过程中可以分为如下几个步骤。首先，网络计算图像$I_m^{(t)}$和I_f通过各自编码器后在嵌入空间中的向量$e_m^{(t)}$和$e_f^{(t)}$之间的距离，这里采用欧氏距离来计算：

$$L_{net}(e_m^{(t)}, e_f^{(t)}) = \left\| e_m^{(t)} - e_f^{(t)} \right\|_2 \tag{3.10}$$

接下来将要计算网络相对于输入$\theta^{(t)}$的梯度$\frac{\partial L_{net}}{\partial \theta^{(t)}}$，这里便是第一次反向计算，但在这一过程中只计算了网络的梯度，并没有更新网络内参数的权重，权重的更新会发生在第二次反向传播中。在[25]中研究者们尝试将$\frac{\partial L_{net}}{\partial \theta^{(t)}}$拆分为旋转和位移两个分量分别进行近似，但因为实际上$\frac{\partial L_{net}}{\partial \theta^{(t)}}$和$\theta^{(t)}$在形状上是一样的，且它要去近似的目标函数$\frac{\partial L_{geo}(\theta^{(t)}, \hat{\theta}^{(t)})}{\partial \theta^{(t)}}$也是一样的，则可以认为这种近似实际上是发生在$\mathrm{SE}(3)$的流形上的。因此本文选择仍然采用在$\mathrm{SE}(3)$的测地线距离来衡量两个梯度间的距离。最终的训练的损失函数如公式 3.11 所示。

$$L = L_{geo}\left(\frac{\partial L_{net}}{\partial \theta^{(t)}}, \frac{\partial L_{geo}(\theta^{(t)}, \hat{\theta}^{(t)})}{\partial \theta^{(t)}}\right) \tag{3.11}$$

在推理阶段，网络的权重被冻结，由于整个框架包括 DRR 投影模块都是可微的，所以

位姿的更新通过采用 PyTorch 提供的 SGD 优化器，根据L_{net}采用梯度下降法去优化输入的初值θ_{ini}。在第t次优化迭代时，如果用$\phi(\cdot)$来表示神经网络这个映射这个过程可以用以下公式表示：

$$\theta^{(t)} = \theta^{(t-1)} - \epsilon \frac{\partial L_{net}(e_m^{(t-1)}, e_f^{(t-1)})}{\phi(I_m^{(t-1)}, I_f)} \frac{\partial \phi(I_m^{(t-1)}, I_f)}{\partial I_m^{(t-1)}} \frac{\partial I_m^{(t-1)}}{\partial P(\theta^{(t-1)}; V)} \frac{\partial P(\theta^{(t-1)}; V)}{\partial \theta^{(t-1)}} \tag{3.12}$$

其中ϵ是当前的优化器的学习率。本文将这个基于深度度量的优化搜索的过程的终止条件设定为当最近十次迭代的搜索位子的方差低于一个界限时搜索终止，即认为此时的优化过程已经收敛至全局最小值附近。

3.4 基于 CMA-ES 的微调

本文的基于强度的微调方法采用 CMA-ES[27]作为优化方法，多尺度的归一化互相关（mNCC）作为相似性度量。CMA-ES 的超参数设置如表 3.1 所示。

表 3.1 CMA-ES 参数设置表

参数名称	参数值
generation_num	15
population_size	100
sigma	0.1
lamda（公式 9）	0.5
upper_bound(degree/mm)	(45, 45, 45, 100, 50, 50)
lower_bound(degree/mm)	(-45, -45, -45, -100, -50, -50)

输入的 mean vector m直接是本文 3.3 小节的模型的预测出的输出结果，整个 CMA-ES 的实现是使用的开源 Python 库 cmaes[49]实现的。整体的算法框架，DRR 投影以及相似性测度计算的部分是在 GPU 上进行的，其余的部分，如演化路径的计算，概率分布参数的更新等则是在 CPU 上完成。为了方便于实现和集成，该步骤仍然采用了与前一节一样的 DRR 生成器。与多数的相关文献一样[36][37]，对于整个优化搜索的过程，早停机制仍然被采用了。与前文中采用的根据最新的几次迭代的标准差来判断优化是否收敛的终止条件不同，这里是否终止是根据相似性度量的最小值来进行判断的，即如果在 5 代之内，记录的相似性度量的最小值没有进一步得到更新，则终止搜索。

3.5 域随机化

因为 CT 图像和 X 光图像具有相同的模态，因此可以通过 CT 扫描生成合理真实的模拟 X 射线图像（DRRs），即完全只使用模拟的 X 射线图像对神经网络进行自监督训练在理论上是可行的，这种方法能够减轻获取大量人工标记数据的负担。但即便是采用现有的最先进的 DRR 生成器[32]进行训练，在真实 X 射线数据上的推理结果仍然存在较大的误差。这表明单纯的 DRR 图像与真实 X 射线图像之间仍然存在较大的域差异（domain gap）。域随机化（domain randomization）是一种用于减小两个域之间的差异的常见方法，它被应用在了机器人[50]以及医学影像[51]许多领域中，在 2D/3D 配准领域也有一些先前的工作[13][52]采用了这一数据增强策略来弥补域间差异。

本文采用的域随机化策略的组成如下：（1）图像平滑：以百分之五十的随机选取核尺寸为3×3或5×5进行平滑操作。（2）噪声注入：向图像中注入均值采样自$(-0.15\cdot max, 0.1\cdot max)$的高斯噪声。（3）重规定化：下界和上界采样间隔分别为$[-0.04\cdot max, 0.02\cdot max]$和$[0.9\cdot max, 1.05\cdot max]$。（4）线性缩放：缩放尺度因子采样自均匀分布$[0.9, 1.05]$。（5）伽马变换：$\gamma$的值从区间$(0.7, 1.3)$中均匀选择。（6）非线性缩放：$a\cdot sin(b\cdot x + c)$，a 和 b 分别从区间$(0.8, 1.1)$中采样，c 从区间$(-0.5, 0.4)$中采样。上述描述中提到的$x$为图像的像素值，$max$为整幅图像的最大强度值。

3.6 本章小结

在本章中，首先介绍了本文所提出的基于自监督学习的 2D/3D 配准的总体流程，接下来分别介绍了该框架的三个子模块：基于深度回归的位姿初始化模块，基于深度度量的位姿搜索优化模块和基于 CMA-ES 的微调。本章通过介绍以上的方法以及它们对应的训练策略和参数设置以及在训练过程中使用的数据增强方法，表述了本文的主要贡献和创新点并且为后续的实验设计与结果进行了铺垫。

第四章 实验设计与结果

本章将介绍围绕本文研究课题和所提出的方法所进行的实验的设计和相对应的结果及其分析。首先会对实验所采用的数据集，数据的预处理流程，提出方法的具体实现以及评价指标分别进行介绍，接下来则会介绍实验环境的设置，最后则是实验得到的结果及其相对于的分析。

4.1 数据集介绍

本课题所使用的数据集均来源自指导老师课题组合作的公司，即邦杰星医疗科技有限公司。其中包含有无配对 X 光的 CT 扫描 465 张（数据集 A），有对应的术中 X 光的 CT 扫描四张(数据集 B)，每份 CT 有对应的正位 Postero-Anterior (PA) 视角的 X 光图像一张，且每份 X 光图像都有一个被预先计算出来的金标准姿态，其中无配对 X 光的 CT 均被选取了腰椎部分作为了感兴趣区域，而有对应的术中 X 光的 CT 扫描则是全身的扫描（full scan CT）。两个数据集的基本信息详见表 4.1。

表 4.1 实验数据集基本信息表

数据集	CT 数量	对应 X 光	人体部位	分辨率
A	465	-	腰椎	0.348~0.711
B	5	5	腰椎	0.779~1.250

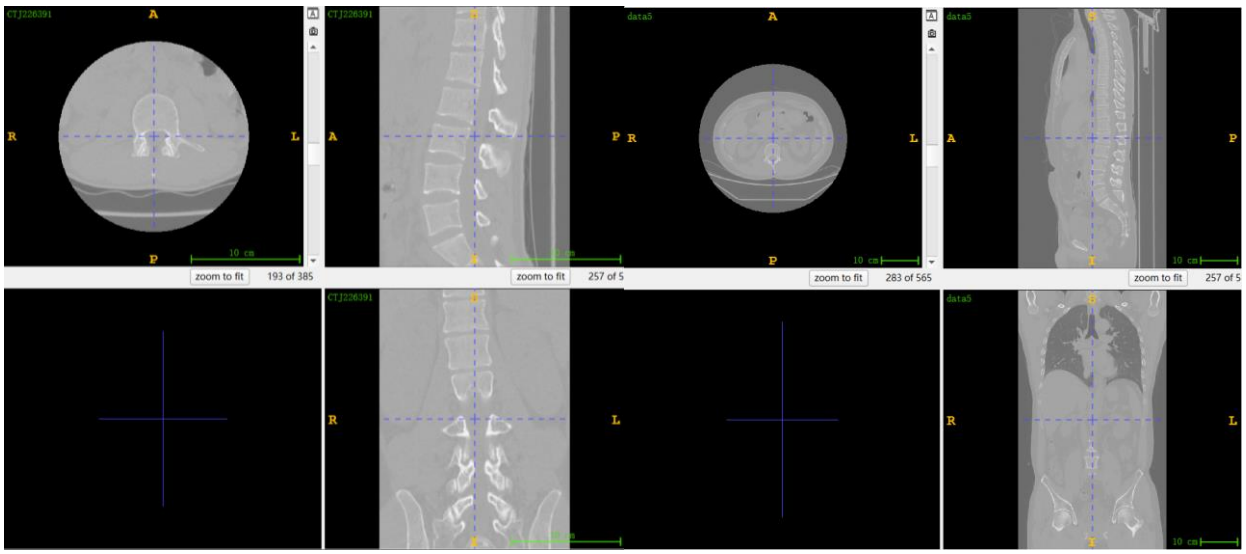

图 4-1 数据集中的 CT 扫描可视化

每份 X 光数据都采集自一台 Perlove PLX118F C-Arm, 它的探测板成像尺寸为1024×1024, 图像的像素间距是 0.199 mm/pixel, 光源点至探测板的距离是 1012 mm，且图像对应的人体位姿都是经过合作企业的工程师采用基于优化的方法进行配准之后再手动调试并根据融合图验证得到的。

4.2 数据预处理

因为数据集 A 的数据已经划定了感兴趣区域（ROI），所以针对图像的中心位置不再做任何调整。首先将图像的各向异性体素间距统一至 0.5mm/voxel，在这一过程中所使用的插值方法为三线性插值，接着采用拼接和裁剪的操作，在保留图像中心不改变的情况下将图像尺寸统一至$512 \times 512 \times 512$，最后再降采样至$256 \times 256 \times 256$。而相较而言，针对数据集 B 的预处理流程则相对更加复杂。首先手动确立一个在 CT 图像中的腰椎部位的中心点，并将该点作为图像的中心（即将 CT 图像的 ROI 选取为以腰椎为中心），再统一图像的各向异性分辨率，并统一图像尺寸至$512 \times 512 \times 512$最后降采样至$256 \times 256 \times 256$。为了去除软组织对配准的影响，对于数据集 A 和 B 中的 CT 扫描都应用了一种全自动的脊柱定位分割和识别方法[53]，以获取它们对应的只含有腰椎的分割图像V_{seg}。针对 X 光数据，因为术中获得的原始图像测量的是 X 射线的衰减，而生成的 DRR 则恰恰相反，它所描述的是 X 线的吸收，简单的来说就是在原始的 X 光图像中高密度骨组织相对较暗而软组织等低密度的人体部位则更为明亮。在 DRR 图像中人体的椎骨部分对应的区域为高亮，其余的是软组织、脂肪或者空气等成分的区域则相对灰暗。图 4-2 给出了一个原始 X 光图像与 DRR 图像之间差异性的可视化描述说明。

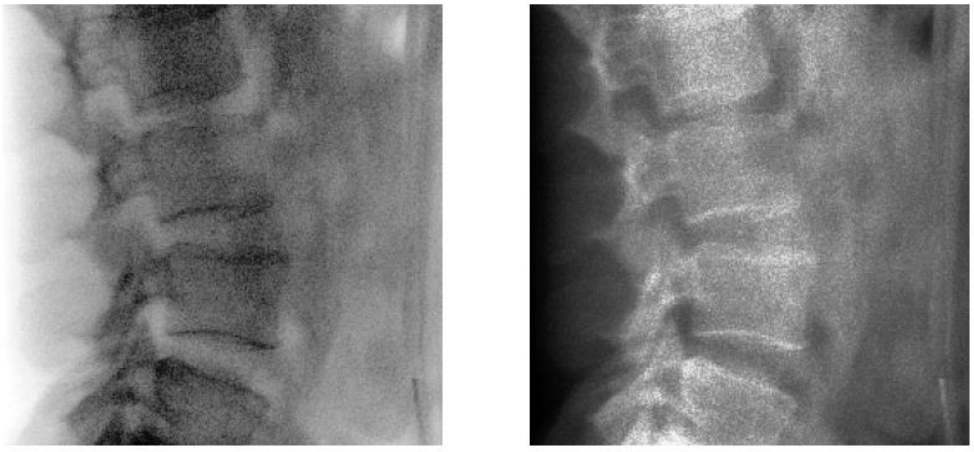

图 4-2 原始 X 光图像（左）与 DRR 图像示意图（右）

假设射线的初始能量为I_0，一旦达到p（即在$\alpha = 1$），这一过程的衰减的能量可以由比尔-兰伯定律（Beer-Lambert law）[54]表示（公式 2.2）。为了使得真实的 X 光图像与用于训练和测试的 DRR 图像灰度分布近似，因而需要对术中的 X 光图像上的每一个点$I(p)$进行在对数域的灰度反转的预处理操作，具体的反转操作可以表示为如下的式子：

$$I_u(p) = 1 - \frac{log(I(p))}{log(I_0)} \tag{4.1}$$

其中射线的初始能量I_0在本实验中是在所有 X 光图像中的灰度最大值，因为部分数据在成像过程中存在过曝，所以实际测量得到的图像强度最大值为 65535.0，这里的对数化操作是为了使翻转的图像更加服从高斯分布，具体实现时为了避免出现负数采用了$\ln(1+p)$来进行对数化。反转后的图像（如图 4-3 所示）将被降采样至尺寸为256×256。

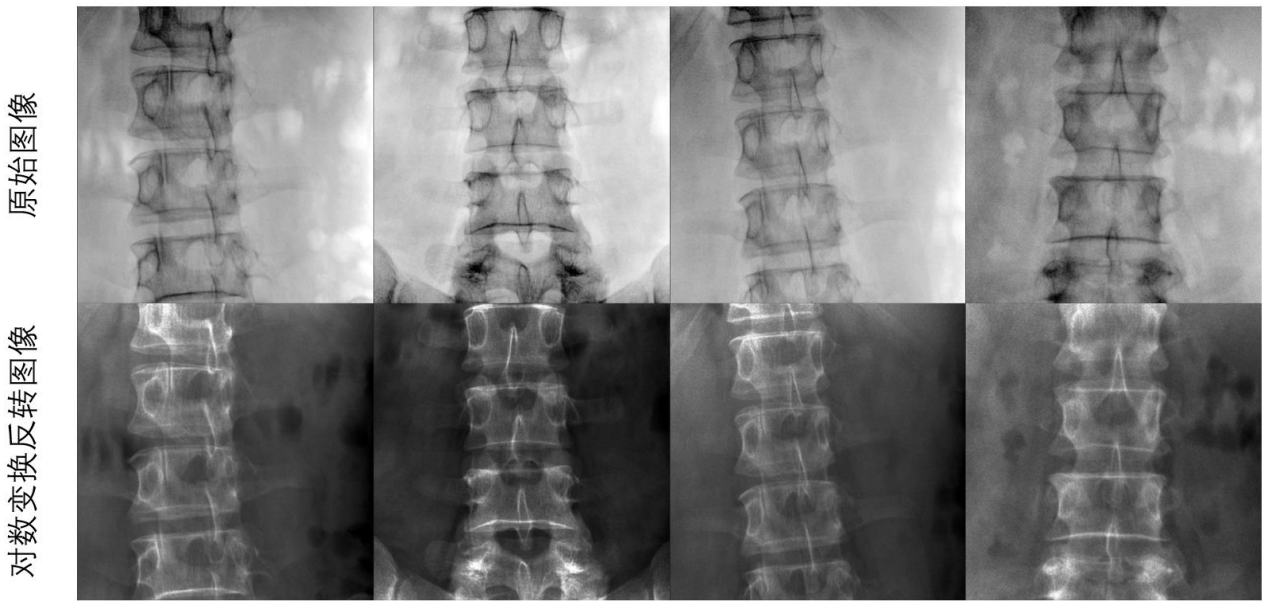

图 4-3 原始 X 光图像（上）与经过预处理之后的 X 光图像（下）

4.3 实现细节

本研究采用自监督的方法对所提出的刚体变换姿态初始化网络（RTPI）和基于深度度量的位姿搜索优化模块进行训练。这主要是由于 2D/3D 的脊柱图像配准任务目前没有开源的数据集且合作公司提供的成对的 X 光与 CT 数据（即数据集 B）太少，无法支撑全监督学习的训练。因此自监督的训练策略被采用，具体来说，在训练的每一次迭代中都随机从

数据集 A 中抽取一张经过预处理的 CT 扫描，并在一个预先设定的分布中采样得到模拟的位姿，再使用当前最先进的 X 光合成框架 DeepDRR[32]生成选取的 CT 在该姿态下对应的模拟 DRR 图像。接下来本小节将分别介绍刚体变换姿态初始化网络（RTPI）和基于深度度量的位姿搜索优化模块是如何在这种策略下训练的。

因为 RTPI 网络的输入为 CT 扫描的分割V_{seg}和配准的目标图像I_t，因此只要采样并生成一张 DRR 图像训练。在采样位姿的过程中，对于平面内（X 和 Y）的方向，本文使用的平移范围为 (-50, 50) mm，而对于深度 (Z) 方向，本文使用的平移范围为 (-100, 100) mm。而对于所有三个轴的旋转方向的参数的的采样范围都是(-20, 20) 度，采样的分布均为均匀分布。这个网络使用 SGD 优化器进行训练，并且采用了循环学习率的调度器，参数设置为每 100 步 学习率在 0.01 到 0.001 之间，动量为 0.9，整个训练进行了 50k 次迭代，batchsize 为 2。公式 3.4 中的参数被分别设置为$\alpha = 1$，$\beta = 1$且正则化参数λ的大小被设为 0.01。

对于基于深度度量的位姿搜索优化模块，因为该模块需要一个初始位姿θ_{ini}以及一个目标 DRR 图像I_t和被分割后的V_{seg}，从而需要有一对位姿被采样。目标图像的位姿θ_t的采样分布与训练 RTPI 模型的一致，θ_{ini}则是在θ_t的附近以正态分布的形式采样随机生成，具体来说，三个旋转方向的采样分布均为 N(0, 15) 度，而三个平移参数的采样范围为 N(0, 20) 毫米。该网络使用 SGD 优化器进行训练并且同样采用了循环学习率的调度器，参数设置为每 100 步学习率在 0.001 到 0.0001 之间，动量为 0.9，整个训练进行了 200k 次迭代。因为训练该模型所需要占用的显存较大，训练的 batchsize 只能被设置为 1，所以梯度累加策略被采用，在训练中梯度被累加 4 次才会更新模型内部的参数。公式 3.11 的测地线距离采用的是由[55]实现的 left canonical 规则下的黎曼度量。在推理阶段，优化过程采用了 SGD 优化器，初始学习率为 0.005，动量为 0.9，同样也被采用的 CyclicLR 调度器的参数设置为每 100 步 学习率在 0.006 到 0.004 之间，搜索的终止条件为近五十次迭代的方差小于 0.001。

整个实验的代码均采用 PyTorch 实现，实验环境为一台配有四张 NVIDIA GeForce RTX 3090GPU 的服务器（由于计算资源限制实验只在其中的一张卡上完成），其配有 20TB 的磁盘空间和多块 2.3-GHz quad-core Intel Core i7 CPU。

4.4 评价指标

在本研究中，两种评价指标将被报告，位姿的误差（pose error）和 2D/3D 任务最为常

见的标准化的评价指标平均目标配准误差（mTRE）[56]。位姿误差是用来计算预测的位姿$\theta \in \mathfrak{se}(3)$和 ground truth $\theta_t \in \mathfrak{se}(3)$之间的差异，为了便于观察，位姿误差结果$|\theta - \theta_t|$被拆分为了旋转和位移两部分，单位分别为角度和毫米。

平均目标配准误差平均目标配准误差度量计算估计姿态和 ground truth 姿态下相应解剖标志之间的平均距离。 假设有一个由K个解剖标志组成的三维的点集Φ，mTRE 可以表示为：

$$\mathrm{mTRE}(\theta, \theta_t) = \frac{1}{K}\sum_{v\in\Phi}\|\theta \circ v - \theta_t \circ v\|_2 \tag{4.2}$$

$$\theta \circ v = exp(\theta)[v, 1]^T \tag{4.3}$$

因为在本课题中使用的三维 CT 数据并没有对应的解剖标志的人工标注，所以分割出的整个腰椎椎骨掩膜都被看作为三维点集Φ。

4.5 实验设置

DRR 生成模块的固有参数（相机参数）被设置来模拟用于采集真实 X 光图像的成像设备，一台 Perlove PLX118F C 型臂移动平台。该设备的成像像素间距为 0.19959mm/pixel，成像尺寸大小为1024×1024，为光源点到探测板的距离为 1011.7 mm。在模拟该设备的成像环境的时候，CT 被放置在成像空间的中心位置，为了降低 GPU 的显存消耗，生成的 DRR 的图像尺寸被减小至了256×256，对应的像素间距也被增大了四倍。

实验讲座 4.1 小节中提到的 A 和 B 两个数据集上进行，其中数据集 A 被用于训练，验证和测试，含有真实 X 光的数据集 B 仅被用作于测试。数据集 A 中随机选取 371 张 CT 作为训练集，47 张作为验证集，47 张作为测试集。验证集与测试集都被随机选取各自生成了一千张 DRR 图像，生成图像的位姿分布与 4.3 小节中的目标图像的位姿生成范围θ_t一致。在数据集 B 中的每张 X 光图像都进行了 50 次实验，最后统计计算它们的平均值。

本文的基线方法包括一种基于学习的 2D/3D 配准方法 DeepReg[25]以及一种基于强度的 2D/3D 配准方法（CMA-ES）[37],为了保证实验的公平性所有的基准方法都是在单一尺度的图像数据上进行了，多分辨率的框架虽然能够进一步优化配准的性能，但考虑到当前计算资源的局限性本研究只在单尺度场景下验证。实验将分别报告本文提出方法以及基线方法在模拟数据以及真实 X 光数据上的表现，报告中采用的评估指标包括位姿的旋转和平移分量的误差以及平均目标配准误差（mTRE）。此外也将验证提出的域随机化方法是否能有效弥补真实 X 光以及模拟数据的域差异。

4.6 实验结果

通过在模拟和真实数据集上进行了基线实验与本文提出框架的实验，我们得到了以下的结果，如表 4.2 和表 4.3 所示。表中所采用的评价指标为在本文第 4.4 小节所介绍的 mTRE 以及位姿在旋转和位移两个分量上的误差，这些评价指标的值越小，表明方法所得到的配准结果与真值之间的误差越小。整个实验中的对比方法除了包括本文提出方法与 4.5 小节中提到的基线方法的对比之外，还包括对提出的三阶段的 2D/3D 配准框架的消融实验，包括去掉最后的基于 CMA-ES 的微调方法（RTPI+FineReg），去掉基于深度度量的位姿搜索优化（RTPI+CMA-ES），以及仅使用基于深度回归的位姿初始化模块的配准（RTPI）。

表 4.2 对比实验各方法在模拟数据上的实验结果

实验方法	**mTRE↓**	**旋转（°）↓**		**位移（mm）↓**	
		mean±std	**median**	**mean±std**	**median**
Random guess	259.2±116.8	33.9±13.3	33.3	63.3±24.5	60.1
CMA-ES	187.6±104.4	25.8±11.3	25.0	74.8±28.0	72.8
DeepReg	179.8±117.5	25.5±15.4,	24.9	54.0±26.2	51.4
RTPI	196.5±61.3	25.8±14.1	25.8	69.6±16.9	73.9
RTPI+CMA-ES	92.9±48.1	11.9±5.2	11.3	23.7±10.3	22.9
RTPI+FineReg	145.6±45.5	19.26±7.1	20.4	30.2±12.4	29.5
Ours	**47.0±90.7**	**3.1±5.9**	**0.7**	**8.2±10.8**	**4.3**

实验结果表明，相较于基于学习的基线方法 DeepReg 和基于强度的 CMA-ES，本文提出的框架在实验中表现出了更高的精度，具体来说，因为本文实验的应用场景为单视角的 2D/3D 配准，所以配准图像天生的对于深度方向上的位移不够敏感，这也可以从 CMA-ES 相较于 Random guess 虽然在旋转分量上的误差以及 mTRE 都有一定的减少，而位移误差变化不明显可以看材。而相较于单独只采用 CMA-ES 的基准方法，基于学习的能够的 DeepReg 明显能够轻易地忽略掉诸多局部极值的干扰，将位姿优化至全局最优附近，这表现是基于深度度量的方法它的度量函数更加平滑，从而对于局部极值不敏感，适用于全局搜索。而即使是与基准方法中在精度上表现最为出色的 DeepReg 相比，本文提出的框架的

mTRE 的均值提高了接近一倍，且在旋转和位移两个分量上的提升也十分明显。这既体现出了本文提出的三阶段框架具有非常良好的全局搜索能力，同时与 DeepReg 相比，本框架在精度上有着明显的优势。因为实验的位姿差异实际上与真实的手术中患者可能出现的位姿变化相比更大，即本实验的模拟数据大多都近似于手术中的极端情况，所以传统的基于强度的方法很容易直接陷入了局部最小值中，导致配准的结果表现不佳。而本文提出的框架能够使得搜索稳定的收敛至全局最优值附近，这体现出该框架有着很好的鲁棒性和稳定性。

表 4.3 对比实验各方法在真实数据上的实验结果

实验方法	**mTRE↓**	**旋转（°）↓**		**位移（mm）↓**	
		mean±std	**median**	**mean±std**	**median**
Random guess	176.8±55.3	29.7±13.1	27.5	56.4±21.5	55.2
CMA-ES	85.3±45.8	22.7±9.6	18.5	51.0±27.6	22.4
DeepReg	95.3±50.6	22.0±5.4	21.7	24.0±9.6	23.8
RTPI	118.4±37.0	21.6±9.3	15.6	53.6±29.6	49.4
RTPI+CMA-ES	49.9±21.6	11.0±9.3	9.6	19.2±15.4	8.1
RTPI+FineReg	94.9±30.1	12.3±6.1	11.2	22.0±7.7	21.4
Ours	**47.1±77.4**	**4.5±7.9**	**0.3**	**13.6±21.2**	**1.5**

从消融实验的结果来看，单独使用 RTPI 模块的精度较差，但它仍然强于 Random guess 的结果，这表明本文提出的基于深度回归的位姿初始化方法有着明显的优势。通过 RTPI+CMA-ES 与完整的三阶段的结果对比表明，基于深度度量的位姿优化模块具有将位姿搜素至了全局最优附近的能力，在不使用该模块的情况下直接使用 CMA-ES 进行微调的配准结果精度明显相对较差，这主要可以归咎于传统的基于强度的配准方法的捕捉范围相对较小，所以存在一些配准样例陷入的局部极值中，导致结果不够理想。通过 RTPI+FineReg 与完整的三阶段的结果对比表明，基于 CMA-ES 的微调步骤有助于保证配准结果的精度，这主要是因为 FineReg 模块虽然具有非常好的全局搜索能力，但它的函数过于平滑所以会在全局最优解的附近过早收敛，从而很难得到一个精度较高的结果，而 CMA-ES 则很好的

保证了结果的精度。基于上述的结果分析，可以看出本文所提出的这种基于自监督学习的三阶段的 2D/3D 配准框架每个模块的有效性，也验证了这种基于自监督学习的方法在 2D/3D 配准领域的前景。这种全自动的 2D/3D 配准方法不仅具有一定的应用潜力，更为实际应用提供了一种可靠的解决方案。

从整体运行的时间上来看，DeepReg 的配准时间大约为 25 秒左右，CMA-ES 则大多需要 40-50 秒，本文提出的框架相对时间较长，在不更改第三阶段微调参数的情况下整个配准框架需要接近一分钟的时间完成配准，在应用了提前停止条件之后运行时间可以减少到 40 秒，但相对的配准的精度也会有一定的下降。实验中汇报的结果均为未采用早停的 CMA-ES 的微调结果。通过对比本文提出框架在模拟数据和真实数据上的实验结果，本文提出的采用自监督学习的训练策略的三阶段的2D/3D配准框架在真实数据上性能并没有出现明显的衰减，这也展示了自监督学习在 2D/3D 配准这个缺乏公开人工标记数据并且人工标记数据困难的领域的应用潜力。

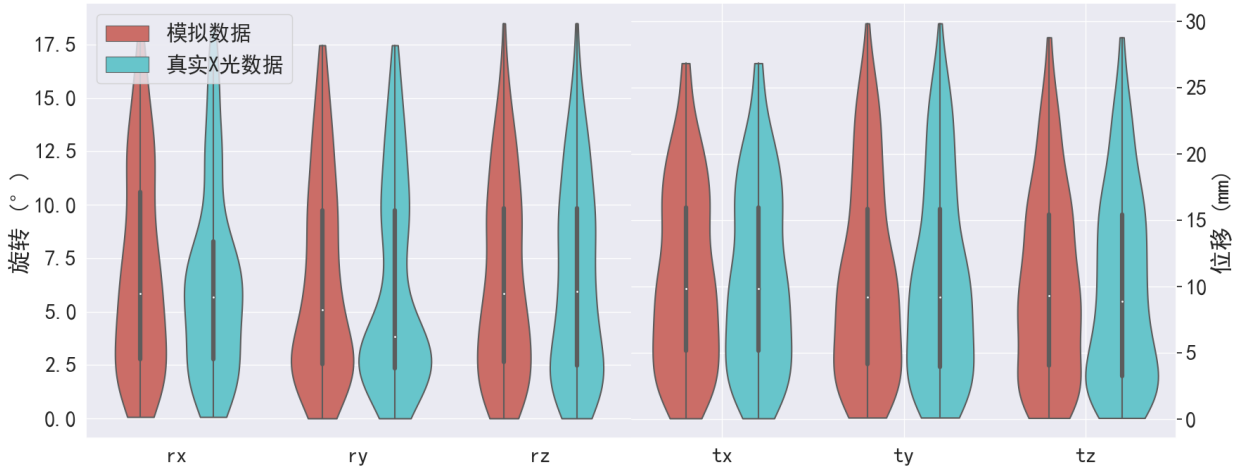

图 4-4 本文提出的 FineReg 模块在模拟数据和真实 X 光数据上配准得到的旋转与位移误差

为验证提出的域随机化方法有效的减小了真实数据和模拟数据两个域之间的差异，在本文提出的采用该策略的基于深度度量的位姿搜索优化模块上进行了比较实验。图 4-4 展示了本文提出的基于深度度量的位姿搜索优化模块（FineReg）在模拟数据和真实 X 光数据上的旋转与位移误差的可视化结果。在该实验中，为了保证公平性，实验统一了真实数据和模拟数据实验中的输入的待优化的位姿，并且模拟数据的真值也与真实数据的真值一致。从图中可以看出，除了在旋转上 X 和 Y 方向上的两个分量外，采用本文提出的域随机

化方法（详见 3.5 小节）训练得到的模型在真实的 X 光数据以及模拟数据上进行配准的效果基本不存在较大的差别。这也验证了提出的域随机化策略能够有效的减小真实 X 光数据与模拟的合成 DRR 数据之间的域差异。

4.7 本章小结

本章主要讲述本文研究的实验设计与结果。首先介绍了实验选取的数据集的基本情况（图像数量，是否带有对于真实 X 光数据等），再而介绍了数据预处理过程（对 CT 扫描的预处理和椎骨的分割以及对 X 光图像的对数域灰度反转等预处理等），接着介绍了模型的实现以及训练的细节，其次介绍了本文实验采取的评价指标（旋转、平移误差以及平均目标配准误差），同时介绍了实验环境的设置以及对比方法的选取来源，最后给出本文的实验结果，并且进行了结果分析。

通过本章的实验结果来看，本文提出的基于自监督学习的三阶段的 2D/3D 配准框架能够取得良好的表现；同时，该框架能够在真实数据集上具有良好的稳定性和鲁棒性。

第五章 总结与展望

5.1 工作总结

本文通过调研 2D/3D 配准任务的目前研究现状，发现现有大多方法都采用在真实数据上进行训练，而本任务既缺乏开源的真实数据集且获取大量人工标记数据的是要耗费大量人力物力的，因此采用在线生成的模拟数据的自监督训练机制在本任务中是具有研究价值的。同时，在目前的文献中，缺乏一种简洁的位姿初始化方法。为此，本文的主要研究内容是设计一个基于自监督学习的 2D/3D 配准框架，期望实现无需人工干预的全自动 2D/3D 配准，且避免了大量人工标记数据的需求负担。

本文首先设计了一种能够快速粗略的预测术中 X 光图像的位姿方法用于获取后续优化搜索位姿的初值。本文采用一种基于深度回归的方法来完成位姿的初始化，该模块首先致力于解决输入的 2D 术中 X 光图像与 3D 的术前 CT 图像在特征融合前完成信息融合。此外无监督的相似性函数和有监督的标签间的 MSE 损失都被应用于训练。接下来本文提出一种通过使用神经网络近似在SE(3)空间中流形上的测地线距离的深度度量方法来优化搜索 2D/3D 的位姿，该方法在全局上函数的形状更为平滑更加近似于凸的，从而增大了配准的捕捉范围。最后本文遵循传统的基于强度的方法的思路设计了一种基于 CMA-ES 的微调模块，用于保证结果的精度和框架整体的鲁棒性。

本文通过调研很多 2D/3D 配准领域的研究，同时考虑了计算的可行性、时效性等，最终确定以平均目标配准误差和位姿的旋转和平移分量的误差评价指标作为研究对象。

本文实验设计选取了由公司提供的私有数据集，包括不含有配对的 X 光图像的单独 CT 数据和包含对应的真实 X 光图像的 CT 数据集作为实验数据集，同时选取了基于学习和基于强度的 2D/3D 配准方法作为实验的基线方法，在相同环境下进行对比实验。

本文的实验结果表明了本文提出的方法在精度和鲁棒性上有着明显的优势，并且自监督的训练方法在真实数据上的表现展示了自监督学习在 2D/3D 配准这个缺乏公开人工标记数据并且人工标记数据困难的领域的应用潜力。

通过本文研究，基本可以表明自监督学习在 2D/3D 配准领域具有重要的应用价值，能够有效解决数据人工标记困难的问题。且提出的三阶段的配准框架为全自动的 2D/3D 配准提供了一种方案和范式，昭示了将基于学习和传统基于优化的方法相结合会是一种更为合理的解决方案。

5.2 工作展望

本文所提出的方法虽然在精度和鲁棒性上展现出了优异的表现，但在运行时间上相对较长，拥有较大的优化空间，未来可以在 GPU 并行加速，以及各个模块各自的终止条件以及相互的衔接方式上进行改进，如当初始化后的位姿已经接近全局最优解时，可以直接使用第三阶段的基于 CMA-ES 的方法来微调，从而减少框架的运行时间。此外位姿初始化模块的性能和可解释性也存在一定的优化空间，未来可以尝试采用在SE(3)中的测地线距离作为损失，并尝试对不同维度的特征提出更为合理的融合方法，如引入单视角重建作为约束以增强过程的可解释性等。

此外，由于现有数据的局限性，本文仅报告了在腰椎数据上的配准表现，未来希望能够在更多的部位如胸椎颈椎数据上进行实验，以全面的评估方法的性能。也希望本文的方法能为其他的研究者提供一些启发和灵感。

参考文献

[1] LaRose D, Bayouth J, Kanade T. Transgraph: Interactive intensity-based 2D/3D registration of X-ray and CT data[C]//Medical Imaging 2000: Image Processing. SPIE, 2000, 3979: 385-396.

[2] Klima O, Kleparnik P, Spanel M, et al. Intensity-based femoral atlas 2D/3D registration using Levenberg-Marquardt optimisation[C]//Medical Imaging 2016: Biomedical Applications in Molecular, Structural, and Functional Imaging. SPIE, 2016, 9788: 113-124.

[3] Miao S, Wang Z J, Zheng Y, et al. Real-time 2D/3D registration via CNN regression[C]//2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI). IEEE, 2016: 1430-1434.

[4] Miao S, Wang Z J, Liao R. A CNN regression approach for real-time 2D/3D registration[J]. IEEE Transactions on Medical Imaging, 2016, 35(5): 1352-1363.

[5] Grupp R B, Unberath M, Gao C, et al. Automatic annotation of hip anatomy in fluoroscopy for robust and efficient 2D/3D registration[J]. International Journal of Computer Assisted Radiology and Surgery, 2020, 15: 759-769.

[6] Grupp R B, Armand M, Taylor R H. Patch-based image similarity for intraoperative 2D/3D pelvis registration during periacetabular osteotomy[C]//OR 2.0 Context-Aware Operating Theaters, Computer Assisted Robotic Endoscopy, Clinical Image-Based Procedures, and Skin Image Analysis: First International Workshop, OR 2.0 2018, 5th International Workshop, CARE 2018, 7th International Workshop, CLIP 2018, Third International Workshop, ISIC 2018, Held in Conjunction with MICCAI 2018, 2018: 153-163.

[7] Van Der Bom I M J, Klein S, Staring M, et al. Evaluation of optimization methods for intensity-based 2D-3D registration in x-ray guided interventions[C]//Medical Imaging 2011: Image Processing. SPIE, 2011, 7962: 657-671.

[8] Penney G P, Weese J, Little J A, et al. A comparison of similarity measures for use in 2-D-3-D medical image registration[J]. IEEE Transactions on Medical Imaging, 1998, 17(4): 586-595.

[9] Song J, Yang K, Zhang Z, et al. Iterative PnP and its application in 3D-2D vascular image registration for robot navigation[J]. arXiv preprint arXiv:2310.12551, 2023.

[10] Sun Y, Zhang H, Chen X, et al. Fast X-ray/CT image registration based on perspective projection triangular features[J]. Computerized Medical Imaging and Graphics, 2024, 112: 102334.

[11] Miao S, Piat S, Fischer P, et al. Dilated FCN for multi-agent 2D/3D medical image

registration[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2018, 32(1).

[12] Esteban J, Grimm M, Unberath M, et al. Towards fully automatic X-ray to CT registration[C]//Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, 2019: 631-639.

[13] Grimm M, Esteban J, Unberath M, et al. Pose-dependent weights and domain randomization for fully automatic X-ray to CT registration[J]. IEEE Transactions on Medical Imaging, 2021, 40(9): 2221-2232.

[14] Markova V, Ronchetti M, Wein W, et al. Global multi-modal 2D/3D registration via local descriptors learning[C]//Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, 2022: 269-279.

[15] Shrestha P, Xie C, Shishido H, et al. X-ray to ct rigid registration using scene coordinate regression[C]//Medical Image Computing and Computer Assisted Intervention–MICCAI 2023: 26th International Conference, 2023: 781-790.

[16] Peng S, Liu Y, Huang Q, et al. Pvnet: Pixel-wise voting network for 6dof pose estimation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 4561-4570.

[17] Xu Y, Lin K Y, Zhang G, et al. Rnnpose: Recurrent 6-dof object pose refinement with robust correspondence field estimation and pose optimization[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 14880-14890.

[18] Jaganathan S, Kukla M, Wang J, et al. Self-supervised 2d/3d registration for x-ray to ct image fusion[C]//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2023: 2788-2798.

[19] Guo H, Xu X, Xu S, et al. End-to-end ultrasound frame to volume registration[C]//Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, 2021: 56-65.

[20] Chen M, Zhang Z, Gu S, et al. Embedded Feature Similarity Optimization with Specific Parameter Initialization for 2D/3D Medical Image Registration[C]//ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2024: 1521-1525.

[21] Tian L, Lee Y Z, San José Estépar R, et al. LiftReg: limited angle 2D/3D deformable registration[C]//Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, 2022: 207-216.

[22] Gopalakrishnan V, Dey N, Golland P. Intraoperative 2D/3D Image Registration via Differentiable X-ray Rendering[J]. arXiv preprint arXiv:2312.06358, 2023.

[23] Zhang B, Faghihroohi S, Azampour M F, et al. A patient-specific self-supervised model for automatic X-Ray/CT registration[C]//Medical Image Computing and Computer Assisted Intervention–MICCAI 2023: 26th International Conference, 2023: 515-524.

[24] Gao C, Feng A, Liu X, et al. A fully differentiable framework for 2d/3d registration and the projective spatial transformers[J]. IEEE Transactions on Medical Imaging, 2023.

[25] Gao C, Liu X, Gu W, et al. Generalizing spatial transformers to projective geometry with applications to 2D/3D registration[C]//Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, 2020: 329-339.

[26] Chen M, Zhang Z, Gu S, et al. Fully Differentiable Correlation-driven 2D/3D Registration for X-ray to CT Image Fusion[J]. arXiv preprint arXiv:2402.02498, 2024.

[27] Hansen N, Ostermeier A. Completely derandomized self-adaptation in evolution strategies[J]. Evolutionary Computation, 2001, 9(2): 159-195.

[28] Siddon R L. Fast calculation of the exact radiological path for a three-dimensional CT array[J]. Medical Physics, 1985, 12(2): 252-255.

[29] Ruijters D, ter Haar Romeny B M, Suetens P. GPU-accelerated digitally reconstructed radiographs[J]. BioMED, 2008, 8: 431-435.

[30] De Greef M, Crezee J, Van Eijk J C, et al. Accelerated ray tracing for radiotherapy dose calculations on a GPU[J]. Medical Physics, 2009, 36(9Part1): 4095-4102.

[31] Unberath M, Zaech J N, Gao C, et al. Enabling machine learning in X-ray-based procedures via realistic simulation of image formation[J]. International Journal of Computer Assisted Radiology and Surgery, 2019, 14: 1517-1528.

[32] Unberath M, Zaech J N, Lee S C, et al. DeepDRR–a catalyst for machine learning in fluoroscopy-guided procedures[C]//Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, 2018: 98-106.

[33] Gopalakrishnan V, Golland P. Fast auto-differentiable digitally reconstructed radiographs for solving inverse problems in intraoperative imaging[C]//Workshop on Clinical Image-Based Procedures. 2022: 1-11.

[34] Knaan D, Joskowicz L. Effective intensity-based 2D/3D rigid registration between fluoroscopic X-ray and CT[C]// Medical Image Computing and Computer Assisted Intervention–MICCAI 2003: 6th International Conference, 2003: 351-358.

[35] Abumoussa A, Gopalakrishnan V, Succop B, et al. Machine learning for automated and real-time two-dimensional to three-dimensional registration of the spine using a single radiograph[J]. Neurosurgical Focus, 2023, 54(6): E16.

[36] Gong R H, Abolmaesumi P. 2D/3D registration with the CMA-ES method[C]//Medical Imaging 2008: Visualization, Image-Guided Procedures, and Modeling. SPIE, 2008, 6918: 556-564.

[37] Chen M, Li T, Zhang Z, et al. An Optimization-based Baseline for Rigid 2D/3D Registration Applied to Spine Surgical Navigation Using CMA-ES[J]. arXiv preprint arXiv:2402.05642, 2024.

[38] 弥佳,周宇佳,冯前进.基于正交视角 X 线图像重建的 3D/2D 配准方法[J].南方医科大学学报,2023,43(09):1636-1643.

[39] 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.

[40] 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. 2016: 770-778.

[41] Chen X, Meng Y, Zhao Y, et al. Learning unsupervised parameter-specific affine transformation for medical images registration[C]//Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, 2021: 24-34.

[42] Zhang G, Li Z, Li J, et al. Cfnet: Cascade fusion network for dense prediction[J]. arXiv preprint arXiv:2302.06052, 2023.

[43] Liu Y, Wang Y, Wang S, et al. Cbnet: A novel composite backbone network architecture for object detection[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2020, 34(07): 11653-11660.

[44] Zhang H, Wu C, Zhang Z, et al. Resnest: Split-attention networks[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 2736-2746.

[45] Mahendran S, Ali H, Vidal R. 3d pose regression using convolutional neural networks[C]//Proceedings of the IEEE International Conference on Computer Vision Workshops. 2017: 2174-2182.

[46] Hou B, Miolane N, Khanal B, et al. Computing CNN loss and gradients for pose estimation with Riemannian geometry[C]//Medical Image Computing and Computer Assisted Intervention–MICCAI

2018: 21st International Conference, 2018: 756-764.

[47] Salehi S S M, Khan S, Erdogmus D, et al. Real-time deep pose estimation with geodesic loss for image-to-template rigid registration[J]. IEEE Transactions on Medical Imaging, 2018, 38(2): 470-481.

[48] Gu W, Gao C, Grupp R, et al. Extended capture range of rigid 2d/3d registration by estimating riemannian pose gradients[C]//Machine Learning in Medical Imaging: 11th International Workshop, MLMI 2020, Held in Conjunction with MICCAI 2020. Springer International Publishing, 2020: 281-291.

[49] Nomura M, Shibata M. cmaes: A Simple yet Practical Python Library for CMA-ES[J]. arXiv preprint arXiv:2402.01373, 2024.

[50] Tobin J, Fong R, Ray A, et al. Domain randomization for transferring deep neural networks from simulation to the real world[C]//2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2017: 23-30.

[51] Dey N, Abulnaga M, Billot B, et al. AnyStar: Domain randomized universal star-convex 3D instance segmentation[C]//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2024: 7593-7603.

[52] Jaganathan S, Wang J, Borsdorf A, et al. Deep iterative 2d/3d registration[C]//Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, 2021: 383-392.

[53] Zhang S, Chen M, Wu J, et al. Spineclue: Automatic vertebrae identification using contrastive learning and uncertainty estimation[J]. arXiv preprint arXiv:2401.07271, 2024.

[54] Swinehart D F. The beer-lambert law[J]. Journal of Chemical Education, 1962, 39(7): 333.

[55] Miolane N, Guigui N, Le Brigant A, et al. Geomstats: a Python package for Riemannian geometry in machine learning[J]. Journal of Machine Learning Research, 2020, 21(223): 1-9.

[56] Van de Kraats E B, Penney G P, Tomazevic D, et al. Standardized evaluation methodology for 2-D-3-D registration[J]. IEEE Transactions on Medical Imaging, 2005, 24(9): 1177-1189.

附录 A 研究成果

1、基于本论文提出的 2D/3D 脊柱图像配准方法，已发表/在投的学术论文：

[1] **Chen Minheng**, Zhang Zhirun, Gu Shuheng, Kong Youyong(*). Embedded Feature Similarity Optimization with Specific Parameter Initialization for 2D/3D Medical Image Registration. ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 1521-1525.（**CCF-B**）

[2] **Chen Minheng**, Zhang Zhirun, Gu Shuheng, Ge Zhangyang, Kong Youyong(*). Fully Differentiable Correlation-driven 2D/3D Registration for X-ray to CT Image Fusion. 2024 IEEE 21th International Symposium on Biomedical Imaging (ISBI).

[3] Li Tonglong[#], **Chen Minheng**[#], Li Mingying, Zhang Zhirun, Zhang Sheng, Li Chuanyou, Kong Youyong(*). Automatic X-Ray to CT Registration using Embedding Reconstruction and Lite Cross-Attention. (Submitted to MICCAI 2024)

（[#]表示共同第一作者 *表示通讯作者）

2、基于本论文提出的 2D/3D 脊柱图像配准方法获得的软件著作权：

Chen Minheng & Zhang Zhirun, SimpleRegi V1.0, registration number: 2023SR0907053, 03/2023

致　　谢

四年的本科时光如白驹过隙，即将为我的求学之路画上圆满的句号。这一路上，众多的人给予了我无私的帮助和温暖的陪伴，此刻，我满怀感激，首先要向所有陪伴并帮助过我的人致以最诚挚的谢意。

首先，我要感谢我的导师孔佑勇副教授，从论文的选题、算法设计、结果分析到论文的撰写，感谢您始终给予我细心的指导和不懈的支持！您广博的专业知识，严谨的治学精神，精益求精的工作作风，深深感染并激励着我。师者传道授业，在此，向您表示最诚挚的敬意！感谢我的本科导师鲍旭东教授，您用丰富的研究经验和瑰宝级的图像处理知识引领着我不断深入课题研究，您严谨求实的科研态度和自律的生活作风，更是深深地影响了我，成为我人生中宝贵的财富。在这里学生祝您身体健康，退休生活愉快！

感谢张晟，李潼泷，吴俊贤，张姿悦等实验室的研究生学长学姐们，在我进入课题组的两年中你们耐心的指导和帮助是我前进的动力。

感谢相伴四年的室友和同学，和你们在一起的时光是我毕生难忘的美好回忆。

在这里我要特别感谢我的朋友兼高中同学张之润。四年的时光里，我们维持着的长期的饭搭子关系，一同游历烟雨江南，携手合作研究课题，这些珍贵的记忆已深深镌刻在我的心中。值此毕业之际，我衷心祝愿他毕业后身体健康，事业有成，工作顺心，生活中充满欢笑，早日财富自由尽早躺平。

我想衷心感谢我的一群朋友和父母家人。感谢何峤宇、石雅郗、金满家、田城铭、付骏宇、谢雨轩等朋友们，希望我们的友谊能够跨越时间和空间的界限绵延久远，以后即便相距万里，隔着十几个时区也还是相亲相爱一家人。感谢我的父母在我本科期间对我的无条件的支持、信任和爱护，你们是我动力的源泉和精神的支柱！

最后，想把陈奕迅的歌《路……一直都在》中的一句歌词送给自己：

不能后退的时候　不再彷徨的时候

永远向前　路　一直都在