doi:10.12068/j.issn.1005－3026.2026.20240188

ViT与光度约束驱动的胸腔镜场景三维重建算法

张子明[1,2]，韩俊涛[1,2]，李鸣骁[1,2]，覃文军[1,2]
（1.东北大学 医学影像智能计算教育部重点实验室，辽宁 沈阳 110169；
2.东北大学 计算机科学与工程学院，辽宁 沈阳 110169）

摘 要：为了提升胸腔镜手术中可视化水平，辅助医生进行精确操作，提出一种手术场景的三维重建算法．首先，利用视觉变换器（ViT）对胸腔镜图像进行深度估计；然后，结合图像金字塔采用光度约束方法预测相机位姿；最后，通过截断带符号距离函数（TSDF）对场景表面进行重建．胸腔镜图像数据集上的实验结果表明：本文方法在深度估计上实现了绝对误差为0.056、均方根误差为3.843的最佳表现；在位姿估计上旋转误差和平移误差分别为0.071 4和0.002 8，效果均优于其他方法．与现有技术相比，本文方法在三维重建精度和术中可视化指导方面均具有显著优势，为胸腔镜微创手术提供了直观、可靠的术中参考信息，具有潜在临床应用价值．

关 键 词：胸腔镜图像；ViT；深度估计；光度约束；位姿预测；视觉三维重建
中图分类号：U 489 **文献标志码**：A **文章编号**：1005-3026(2026)02-0033-08

3D Reconstruction Algorithm of Thoracoscopic Scenes Driven by ViT and Photometric Constraints

ZHANG Zi-ming[1,2], *HAN Jun-tao*[1,2], *LI Ming-xiao*[1,2], *TAN Wen-jun*[1,2]
(1. Key Laboratory of Intelligent Computing in Medical Image, Ministry of Education, Northeastern University, Shenyang 110169, China; 2. School of Computer Science and Engineering, Northeastern University, Shenyang 110169, China. Corresponding author: TAN Wen-jun. E-mail: tanwenjun@.cse.neu.edu.cn)

Abstract: This study aims to propose a method for three-dimensional (3D) reconstruction of thoracoscopic surgical scenes to enhance intraoperative visualization and assist surgeons in performing precise operations. The method first employs a vision transformer (ViT) to estimate the depth from thoracoscopic images; it then predicts camera poses using a photometric constraint approach combined with an image pyramid; finally, the scene surface is reconstructed using a truncated signed distance function (TSDF). Experimental results on the thoracoscopic image dataset demonstrate that the proposed method achieves the best performance in depth estimation, with an absolute error of 0.056 and a root mean square error of 3.843, and in pose estimation, its rotation and translation errors are 0.071 4 and 0.002 8, respectively, outperforming other methods. Compared with existing techniques, this approach shows significant advantages in 3D reconstruction accuracy and intraoperative visualization guidance, providing intuitive and reliable intraoperative reference information for thoracoscopic minimally invasive surgery, holds potential clinical value.

Key words: thoracoscopic image; ViT network; depth estimation; photometric constraint; pose prediction; visual three-dimensional reconstruction

世界卫生组织国际癌症研究机构发布的2022年全球最新癌症数据显示，肺癌死亡病例远超其他癌症类型，位居癌症死亡类型榜单之首[1]．胸腔镜手术已成为治疗肺部疾病的主流微创手术方法．在手术过程中，胸腔镜虽然可以提供高分辨率、高清晰度的肺部图像，但仅能提供肺部

收稿日期：2024-10-23
基金项目：国家自然科学基金资助项目（62471122）．
作者简介：张子明（1996—），男，河北乐亭人，东北大学博士研究生．
通信作者：覃文军，E-mail：tanwenjun@.cse.neu.edu.cn.

的二维视觉场景,缺乏深度方向的信息.医生通常凭借自身的主观经验确定病灶的空间位置以进行肿瘤切除,很难定量化地作出清晰、直接、具体的客观判断.

计算机辅助手术(CAS)的广泛应用[2]为解决上述问题提供了一种思路,可以设计一种基于计算机视觉技术辅助胸腔镜手术的技术方案.基于这种方案,可以考虑在术中将肺部胸腔镜场景进行三维重建,从而获得深度信息.将重建出的肺部三维表面与术前拍摄的肺部CT图像重建出的肺部模型进行配准、对比,能更方便医生进行术前规划以及在术中定位肺部病灶的具体空间位置.

一些学者将自动驾驶领域常见的基于视觉和电磁跟踪的定位技术在医学胸腔镜领域进行实验,并开展了广泛研究.例如,稠密跟踪与建图(DTAM)[3]是一种用于实时三维重建和跟踪的技术.DTAM的核心思想是通过摄像头捕捉的图像序列,实时构建场景的稠密三维模型,并跟踪相机在模型中的运动.

深度学习方法也被应用于内窥镜三维重建的深度估计部分.Zhou等[4]开发了一种自监督深度估计框架,该框架将深度学习问题转换为视图合成任务.该框架包含1个深度网络和1个单独的位姿网络.为了处理物体运动和图像遮挡等图像边缘信息问题,该框架使用了可预测可解释掩码实现对每个像素点的过滤.许多研究者基于文献[4]的框架加入几何先验来改善模型.

尽管在基于深度学习的深度估计中,学者采用了各种基于先验知识针对不足之处辅以优化的方法,但是对于真实胸腔镜手术场景中采集的胸腔镜图像纹理特征不足的问题以及预测的位姿是否存在"轨迹漂移"并没有展开大量研究.针对上述问题,实现充分提取胸腔镜图像的特征并且对深度学习网络预测的位姿进行优化是可能且有必要的.因此,本文的深度估计与位姿预测方法旨在解决上述问题.

1 材料和方法

1.1 数据集

用于本实验的数据包括以下数据集:①Hamlyn数据集[5].Hamlyn数据集是在外科手术过程中使用内窥镜拍摄的猪腹部场景信息,视频清晰且质量高,并且包含准确的真实深度信息.本文使用的Hamlyn数据集包含14个视频序列,视频分辨率为720像素×288像素.②临床数据集.与医院合作,在使用奥林巴斯公司的内窥镜进行肺部手术时,该内窥镜自带录像功能,在进行手术器械操作前请医生尽可能多角度地拍摄了患者胸腔的情况.其中包括11台完整手术的内窥镜视频,分辨率为1 920像素×1 080像素.

1.2 内窥镜标定

在进行三维重建时,需要利用相机内参来实现像素坐标系与相机坐标系之间的转换,以减小三维重建模型与真实物体之间的差距.因此,相机内参对于三维重建至关重要.相机内参需通过胸腔镜拍摄完整、全方位视角的标定板图像进行标定获取,在肺部胸腔镜手术真实场景下,使用胸腔镜拍摄的12张标定板图像,如图1所示.

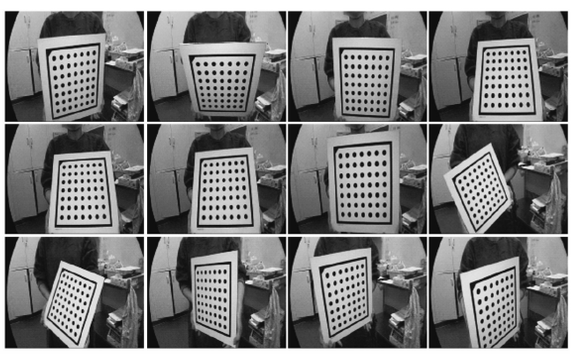

图1 内窥镜标定板采集结果

Fig. 1 Acquisition results of endoscopic calibration plate

将胸腔镜拍摄的标定板图像通过OpenCV进行相机标定,得到相机内部参数$\boldsymbol{K}$如下:

$$\boldsymbol{K}=\begin{bmatrix}1\,035.31 & 0 & 596.96\\ 0 & 1\,035.09 & 520.41\\ 0 & 0 & 1.00\end{bmatrix}. \tag{1}$$

1.3 胸腔镜图像去反光修复

胸腔镜手术真实场景中光照环境不如自然环境稳定,胸腔镜图像序列中会出现前后帧间亮度不一致的情况,特别是一些图像存在像素点过亮的问题,这会导致在深度估计过程中无法有效提取特征进而造成失真,最终影响三维重建的结果.为保持亮度的一致性,本文采用基于反光检测与反光修复的策略来解决图像高亮问题,以得到符合亮度恒常性假设的胸腔镜图像序列.

反光检测的基本思路是首先识别胸腔镜图像中的高光区域,将其视为反光区域,然后利用周围像素信息的平均值对反光区域进行填充,接着对填充区域进行中值滤波以消除噪声,获得平滑的非反光颜色像素[6],最后将高光区域的像素

值与平滑的非反光颜色像素值进行比较,若前者大于后者,则认为该高光区域亮度过高,需要进行去反光处理.

对于反光修复,先将反光区域用反光检测中同样的方式进行填充,然后对该图像进行高斯模糊处理;根据结构相似性指数(SSIM)来评估原始反光图像和经过高斯模糊处理的反光图像之间的相似度,计算出2张图像的权重,距离反光中心近的区域权重要大一些,反之小一些;最后对反光图像和经过高斯模糊处理的反光图像进行加权相加,从而实现反光修复.反光修复效果如图2所示.

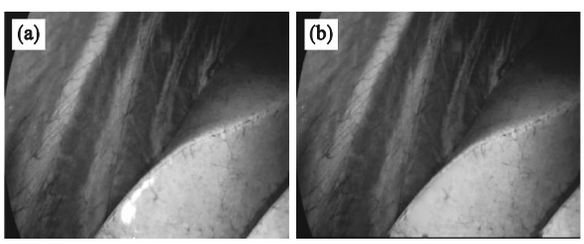

图2　内窥镜去反光结果

Fig. 2　Endoscopic dereflection results

(a)—去反光前;(b)—去反光后.

1.4　基于视觉变换器(ViT)的深度估计方法

1.4.1　特征提取模块

视觉变换器(ViT)是将变换器引入计算机视觉领域的创新模型,本节通过ViT的自注意力机制实现了对胸腔镜图像全局特征的学习和表示.ViT网络特征提取步骤如下:

1)图像分块.为了方便ViT架构能够处理胸腔镜二维图像数据,需要将输入的图像$X\in\mathbf{R}^{H\times W\times C}$重塑为一系列二维图块$X_{\mathrm{p}}\in\mathbf{R}^{N\times(P^2C)}$.其中:$H\times W$为输入图像的分辨率;$N$为图块数量;$P$为重塑后每个图块的尺寸;$C$为通道数.

2)图块嵌入.ViT中所有层均使用恒定维度D的潜在向量,将图块映射到ViT模型的输入维度上,图块的线性投影过程就是图块嵌入.该步骤的目的是将每个图块的像素值展平并投影到低维空间,得到每个图块的嵌入表示.

3)位置嵌入.为每个图块的嵌入表示添加位置编码,以表示图块在图像中的位置信息,然后将经过嵌入和位置编码处理的图块序列输入到ViT模型中.

4)编码器提取特征.首先,ViT经过多头自注意力机制(MSA),每个自注意力头都能学习图块序列中不同位置之间的依赖关系,有助于模型捕捉全局信息和图块信息之间的关联;其次,该前馈神经网络由全连接层和激活函数组成,用于对中间输出进行非线性变换,并进行残差连接和层归一化(LN),这有助于减轻训练深层网络时梯度消失的问题,并加快ViT模型的收敛速度;最后,引入了多层感知机(MLP)结构,用于特征信息的非线性变换.

1.4.2　深度估计

使用Monodepth2[7]网络作为基础的深度估计框架进行深度图的生成,并在Monodepth2网络上采用自动掩码损失策略、计算每个像素点最小重投影损失策略以及全分辨率多尺度采样策略.

自动掩码损失策略.在自动掩码损失中定义了像素点的掩码μ,此掩码是二值化的参数,该掩码在网络前向传播过程中自动计算,计算公式如下:

$$\mu=[\min_{t'}E(I_t,I_{t'\to t})<\min_{t'}E(I_t,I_{t'})]. \tag{2}$$

式中:I_t为目标图像;$I_{t'}$为无畸变的源图像;t'为源图像的时间索引,用来遍历与目标索引t相邻的图像;$I_{t'\to t}$为畸变图像;E为光度误差,具体数值为像素间的L1距离与结构相似性指数之和;$\min_{t'}$为对所有源图像的光度误差取最小值;[ ]为Iverson括号,用于将布尔条件转换为整数值.

每像素点最小重投影误差.根据多个源图像帧计算重投影误差时,通常将重投影误差平均到每个源图像帧中,这可能会导致在目标图像帧中可见但在某些源图像帧中不可见的像素计算出错.最小重投影误差方法仅将目标帧中的每个像素与可见源图像帧中的像素点进行匹配,从而获得更清晰的结果.

全分辨率多尺度采样策略.为了避免在计算解码器每个网络层不同分辨率下图像的光度损失时,低分辨率深度图中大量的纹理不丰富区域可能产生空洞以及视觉伪影,本文对多尺度分辨率计算进行改进,将红绿蓝(RGB)图像的分辨率和用于计算重投影误差的视差图像解耦,采用全分辨率多尺度估计策略,在计算低分辨率图像上的光度损失之前,先将其对应的深度图使用双线性插值方法从中间层上采样到与输入图像相同的分辨率,然后重新投影和采样并计算其在高分辨率下的光度损失.

1.4.3　整体流程和网络结构

输入原始的肺部胸腔镜RGB图像序列,图像经过ViT网络进行特征提取,将ViT提取得到的特征向量以相加的方式并入Monodepth2编码器

提取的肺部胸腔镜RGB图片特征.由于Monodepth2采用了多尺度策略,编码器中生成了4层特征,所以ViT生成的特征向量需要通过卷积操作进行缩放,然后再并入每层提取的特征,重新融合成新的特征图,以上共同构成了新的编码器部分.

对于解码器部分,所有层均为卷积层,卷积核大小均为3×3,步长为1,解码器包含4个相同的反卷积模块,每个反卷积模块的输入为编码器结构中提取的相同尺度的特征图和上1个反卷积模块的输出,将特征图融合后,进行卷积和上采样,使最后1个反卷积模块输出的深度图分辨率与原始输入的RGB图像分辨率相同.为了避免在特征图的边界发生信息丢失或出现伪影等边界效应,在解码器中使用了反射填充来代替零填充,即通过在边界周围复制像素值的方式来填充,而不是简单地填充零值,以更好地保留图像边缘的细节和信息.

1.5 基于光度差异的位姿估计方法

在光度约束法中使用图像金字塔策略进行优化,以提高收敛范围,从而在一定程度上避免落入局部最小值.光度约束法的优势在于进行胸腔镜位姿预测时计算效率高,预测相机位姿的时间相对较短,很适合用于术中胸腔镜场景的实时三维重建.此外,光度约束法有很强的泛化能力,可以应对不同的微创手术环境和场景.

使用基于光度差异的直接法来预测胸腔镜的位姿,估计目标图像关键帧I_k与当前源图像帧I_c之间的相对变换矩阵$\boldsymbol{T}_{ck}$,即胸腔镜的旋转和平移,其中变换矩阵属于李群中的三维特殊欧氏群.位姿预测的本质是不断优化$\boldsymbol{T}_{ck}$,从而使不同图像帧中互相对应的像素点具有最小的光度误差.本方法将预测胸腔镜的运动轨迹定义为Lucas-Kanade[8]框架下的非线性最小二乘问题,以最小化光度误差.

此外,还使用了从粗粒度到细粒度的图像金字塔优化来提高优化算法的收敛范围,具体流程如下:①读取前一帧和当前图像帧;②初始化相机位姿;③遍历图像金字塔每一层,在当前层上计算光度误差,计算雅可比矩阵以估计当前参数变化对目标函数的影响、更新步长以避免参数变化较大;④判断是否遍历到最底层,如果不是最底层,则更新全部相机位姿,反之则只更新旋转部分的位姿;⑤遍历结束后输出优化后的相机位姿.

本节在位姿预测中采用与DTAM[3]方法类似的图像金字塔策略,在图像金字塔中分辨率最低的底层图像(低尺度)上对胸腔镜的旋转运动进行优化.由于该层图像经过较大比例的降采样,其对局部像素细节的敏感度较低,能够在一定程度上平滑胸腔镜运动模糊带来的影响,因此在低尺度层进行旋转优化具有更快的收敛速度和更好的鲁棒性.在图像金字塔由低尺度向高尺度逐层递进的过程中,对每一尺度下的相机位姿进行完整的6自由度(旋转和平移)优化.相机位姿轨迹的局部运动采用李代数表示,并使用高斯牛顿法对相机位姿进行迭代更新.

最终的损失函数$\hat{\psi}$是目标图像关键帧I_k与从源图像帧变换到目标帧$I_{c\to k}$中对应像素的光度误差.

$$\hat{\psi}=\arg\min\sum_{p\in I_k}\min\left(\left\|I_k[p]-I_{c\to k}[p]\right\|_2,\alpha\right). \tag{3}$$

式中:p为像素;[ ]为Iverson括号;α为使残差的二范数趋于饱和的阈值.

从源图像帧I_c到目标图像关键帧I_k的胸腔镜位姿由以下2部分组成:①源图像帧到关键帧的全局变换$\boldsymbol{T}_{ck}^0$,它表示了2帧之间的整体变换;②李代数中局部更新的指数映射,将微小的局部运动变换表示为李代数的参数,然后通过指数映射将其转换回李群的三维特殊欧氏群中,以获得整体的运动变换$\boldsymbol{T}_{ck}$.

$$\boldsymbol{T}_{ck}=\mathrm{e}^{\psi}\boldsymbol{T}_{ck}^0. \tag{4}$$

式中,ψ为位姿优化中李代数的位姿增量.

1.6 腔镜场景三维重建

本节需要先实现胸腔镜数据的三维表面重建.将深度估计获得的胸腔镜图像深度图与原始的胸腔镜图像RGB图组成RGB-深度(RGB-D)图像帧,同时通过位姿预测得到胸腔镜位姿,即胸腔镜外参矩阵,有了以上信息即可以进行胸腔镜视觉场景的三维重建.本节基于截断带符号距离函数(TSDF)[9]进行三维体素的获取与融合,并使用行进立方体算法[10]进行等值面的提取,从而完成胸腔镜图像视觉场景的三维重建.

TSDF算法的流程如下:

1)初始化TSDF三维空间,创建能够包围住场景三维模型的足够大的三维空间,对三维空间等值划分成网格体素.

2)计算当前体素的TSDF值以及权值,则点P_w的TSDF值和权值如下:

$$f(P_w)=(d_p(x)-d_s(P_c))/|\delta|, \tag{5}$$

$$w(P_w)=\cos\theta/d_s(P_c). \tag{6}$$

式中:P_c为点P_w在相机坐标系下的坐标;$d_s(P_c)$为相机光心到P_c的距离;$d_p(x)$为像素点x的深度值;δ为截断因子;δ用来保证TSDF值的二值化;θ为投影光线与体素表面法向量的夹角.

3)体素融合.将多个RGB-D帧投影到TSDF三维空间中进行融合,实际上是将每个图像中的像素对应到三维空间中的体素.通过相机外参,可以将像素映射到三维空间中的位置,然后将这些位置处的信息与体素进行融合.

基于行进立方体算法的胸腔镜视觉场景表面重建具体步骤如下:①MC算法初始化.定义查找表,用于确定每个体素的拓扑结构和生成三角化网格的规则.②遍历体素网格.对于每个体素,根据上述TSDF算法确定等值面在该体素内的位置,根据已知的TSDF值确定哪些体素边界被穿过,从而更新每个体素边界的状态.③顶点生成.对于每个穿过的体素边界,根据线性插值方法,确定等值面与该体素边界交点的具体位置,并生成顶点.④三角化网格生成.根据每个体素的边界状态以及胸腔镜手术影像场景表面穿过的体素边界,使用查找表确定生成三角化网格的拓扑结构.⑤等值面绘制.将生成的三角化网格连接起来,构建出完整的等值面.

2　实验结果与分析

2.1　深度估计实验

2.1.1　实验环境

实验均在Linux操作系统上完成,所需的环境配置信息如下:处理器为12代Intel i7-12700F;显卡为Nvidia RTX 3060;操作系统为Ubuntu 20.04;编程语言为Python 3.7.3;深度学习环境中的程序版本为CUDA 10.2,cuDNN 7.6.5,Pytorch 1.7.0,Torchvision 0.8.0.

将Hamlyn数据集的14个视频进行切帧,共获取到12 630帧图片,把数据集中的无效帧和多余帧去除,并划分训练集和测试集,得到9 016张训练集用于训练深度网络模型,1 705张测试集用于评价性能指标.将肺部胸腔镜手术数据进行视频切帧得到1 892张图片,这些图片经过去反光修复后用于测试并生成深度图.实验使用Adam[11]优化器,其中优化器参数$\beta_1=0.9$,$\beta_2=0.99$.每轮的批次大小为12,训练轮次为20轮.

2.1.2　评价指标

实验采用深度估计领域中常用的误差指标和精度指标,对深度估计模型的性能用以下参数进行评价:绝对相对误差、平方相对误差、均方根误差、对数均方根误差,以及δ值.其中δ的公式如下:

$$\delta=\frac{1}{D}\left\{d\in D\mid \max\left(\frac{d^*}{d},\frac{d}{d^*}\right)<1.25\right\}. \tag{7}$$

式中:d为预测深度;d^*为真实深度;D为计算误差的深度值个数.

2.1.3　与其他方法比较

表1给出了本文方法与Lapdepth[12],IsoNR-SfM[13],Endo-SfM[14]以及AF-SfM[15]在同一数据集下的性能指标对比,误差对比中最好的结果用黑体标出,次之用下划线标出.Lapdepth是目前较先进的有监督学习深度网络,IsoNR-SfM是一种传统的多视图方法.Lapdepth和IsoNR-SfM使用了作者的默认配置进行训练,Endo-SfM和AF-SfM按照2.1.1小节中的参数进行设置.

表1　深度估计实验结果

Table 1　Experiment results of depth estimation

实验方法	绝对相对误差	平方相对误差	均方根误差	对数均方根误差
Lapdepth	0.432	12.182	11.742	0.408
IsoNR-SfM	<u>0.058</u>	**0.285**	6.420	0.084
Endo-SfM	0.068	0.598	5.721	0.096
AF-SfM	0.062	0.433	<u>4.916</u>	<u>0.082</u>
本文方法	**0.056**	<u>0.431</u>	**3.843**	**0.077**

由表1可见,即使在有监督学习的情况下,Lapdepth的表现也明显不如本文方法,因为合成图像和真实图像之间存在领域转移.虽然本文方法的平方相对误差不如IsoNR-SfM,但是本文方法生成了图像的密集深度,而不局限于寻找多视图对应.

另外,本文方法在绝对相对误差、均方根误差和对数均方根误差3个指标上,均达到了最优,尤其是均方根误差与次优的AF-SfM方法相比提高了22%.其中最大的差异是ViT模块的使用,ViT模块让网络能够更专注于重点区域.在胸腔镜环境下,光源只来自胸腔镜和微创切口,这导致了切口处与环境深处的亮度差异过大,而ViT模块可以减轻亮度不合理区域的影响.并且均方根误差指标相比于其他指标,对大误差更加敏感.均方根误差指标的优秀表现,证明了本方法更加稳定.

2.1.4　消融实验

在Hamlyn数据集上,对本文提出的基于ViT

辅助的Monodepth2模型与Monodepth2模型进行了对比实验,性能指标对比结果如表2所示.

表2 深度估计消融实验结果

Table 2 Ablation experiment results on depth estimation

模型	绝对相对误差	平方相对误差	均方根误差	对数均方根误差	$\delta<1.25$	$\delta<1.25^2$	$\delta<1.25^3$
Monodepth2	0.060	0.478	4.105	0.084	0.966	0.997	0.999
ViT辅助Monodepth2	0.056	0.431	3.843	0.077	0.973	0.998	0.999

使用ViT辅助Monodepth2模型进行特征提取与使用Monodepth2模型进行特征提取相比,深度估计的绝对相对误差降低0.004、平方相对误差降低0.047、均方根误差降低0.262、对数均方根误差降低0.007;$\delta<1.25$提高了0.007,$\delta<1.25^2$提高了0.001,$\delta<1.25^3$与原网络保持相同.可见,当使用ViT辅助Monodepth2模型进行特征提取时,深度估计准确度有了一定的提升.

2.2 位姿估计实验

2.2.1 评价指标

在基于光度差异的位姿预测方法中,使用绝对轨迹误差(ATE)对相机位姿进行评价,将旋转和平移拆分开进行对比,以直观地看出相机运动的偏差,并与传统方法进行性能对比.旋转和平移的绝对轨迹均方误差L_{R}和L_{T}如式(8)和式(9)所示:

$$L_{\mathrm{R}}=\sqrt{\frac{1}{N}\sum_{i=1}^{N}(\boldsymbol{R}_{\mathrm{g}}-\boldsymbol{R}_{\mathrm{p}})^2}, \tag{8}$$

$$L_{\mathrm{T}}=\sqrt{\frac{1}{N}\sum_{i=1}^{N}(\boldsymbol{T}_{\mathrm{g}}-\boldsymbol{T}_{\mathrm{p}})^2}. \tag{9}$$

式中:N为相应计算的旋转矩阵和位移矩阵个数;$\boldsymbol{R}_{\mathrm{g}}$为真实相机外部参数中的旋转矩阵;$\boldsymbol{R}_{\mathrm{p}}$为预测的相机外部参数中的旋转矩阵;$\boldsymbol{T}_{\mathrm{g}}$为真实相机外部参数中的平移矩阵;$\boldsymbol{T}_{\mathrm{p}}$为预测的相机外部参数中的平移矩阵.

2.2.2 本文方法与其他方法的比较

对于基于光度差异的位姿预测方法(图像金字塔+光度约束法),采用带有真实相机外参的Hamlyn数据集对光度法及加入图像金字塔优化的光度约束法进行评价.同时,为证明本文方法相比传统方法具有更好的预测效果,又与点到点的迭代最近点(ICP)算法[16]以及点到平面的ICP[17]算法进行了对比,2种ICP算法均在Open3D软件[18]下实现.将这些方法分别在相同的图像序列上进行位姿估计,并计算各方法旋转和平移预测结果的误差,最终结果如表3所示,最好的结果由黑体标出.

对比光度约束法,在本文方法中使用图像金字塔优化后,相机外参矩阵的旋转均方误差减小了0.003 0,平移均方误差减小了0.000 1,旋转均方误差性能提升较大的原因可能是本方法中使用了粗粒度的图像金字塔单独优化了相机的旋转,从而使高斯-牛顿优化算法的收敛更快,配合细粒度的图像金字塔优化6自由度位姿,进一步提升了预测相机平移运动的准确性.

表3 位姿估计实验结果

Table 3 Experiment results of pose estimation

实验方法	L_{R}	L_{T}
点到点ICP算法	0.083 4	0.004 6
点到平面ICP算法	0.081 1	0.003 7
光度约束法	0.074 4	0.002 9
图像金字塔优化的光度法	**0.071 4**	**0.002 8**

2.2.3 消融实验

图像金字塔优化的光度约束法与光度约束法预测的位姿轨迹与真实位姿轨迹对比如图3所示.图3a为图像金字塔优化的光度约束法位姿预测结果;图3b为光度约束法位姿预测结果.由此看出,加入图像金字塔优化后,位姿轨迹与真实轨迹更加贴近.由于ICP算法属于传统基于特征点的位姿预测方法,较难处理内窥镜影像的纹理信息,而光度约束法属于直接法,直接对像素信息进行计算,因此图像金字塔优化的光度约束法与光度约束法的性能均好于ICP算法.此外,在点到点ICP算法中是将目标点云中的每个点与参考点云中的最近邻点进行匹配,而点到平面ICP算法中还考虑了目标点与最近邻点所在平面的距离,对于内窥镜场景下纹理不足或含有噪声的数据,相比点到点ICP算法能够产生更准确的位姿预测结果.

2.3 三维重建结果

在临床肺部胸腔镜真实手术场景中采集的2个胸腔镜手术图像序列上进行了视觉三维重建.由于第2个图像序列场景过暗,多张图像重建的效果不利于观察,因此对图像序列2的第1帧图像又进行了单帧图像的三维重建.2组图像序列的三维重建结果分别如图4和图5所示.

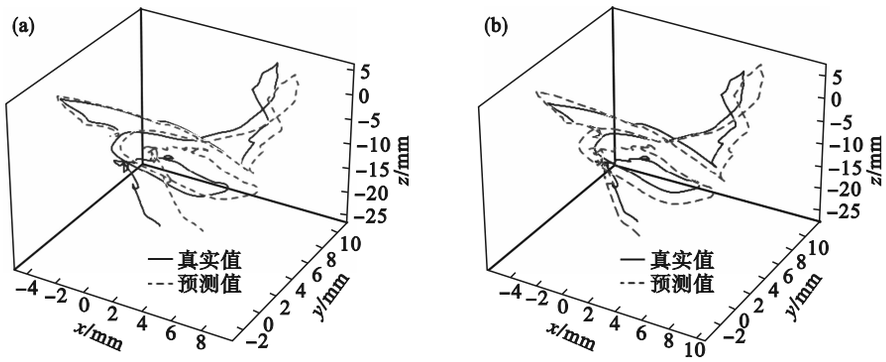

图3　位姿估计消融实验结果

Fig. 3　Ablation experiment results of pose estimation

(a)—图像金字塔优化的光度约束法;(b)—光度约束法.

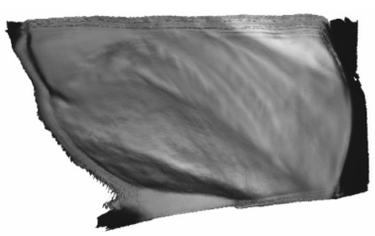

图4　序列1的多帧三维重建结果

Fig. 4　3D reconstruction results of multiple images for sequence 1

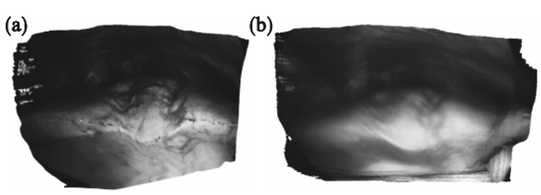

图5　序列2的三维重建结果

Fig. 5　3D reconstruction results for sequence 2

(a)—第1帧重建;(b)—全部序列重建.

为了定量验证重建点云精度,选取了1组带有手术器械的腔镜图像进行重建,重建结果如图6所示.以手术器械的长度作为基准,验证重建点云的精度.使用MeshLab软件测量了点云上方器械从尖端到末端的长度为4.086 58 cm,下方器械的长度为3.424 66 cm.与实际腔镜图像中器械长度的误差在10%以内,证明了本文方法重建点云的准确性.

图6　带手术器械的三维重建结果

Fig. 6　3D reconstruction results with surgical instruments

3　结　　语

本文提出了一种适用于胸腔镜手术场景的三维重建方法,在深度估计和相机位姿预测中分别引入视觉变换器网络和结合图像金字塔的光度约束策略,并利用截断带符号距离函数对场景表面进行重建.实验结果表明,本方法在深度估计上实现了绝对误差0.056、均方根误差3.843,位姿估计的旋转均方误差0.071 4、平移的绝对轨迹误差0.002 8,均优于现有对比方法,并生成连续、稳定的三维场景表示.该方法不仅显著提升了胸腔镜手术中可视化的直观性和准确性,也为微创手术提供了可靠的术中导航信息,具有重要的临床应用潜力.

参考文献:

[1] World Health Organization. Global cancer burden growing, amidst mounting need for services[EB/OL](2024-02-01)[2024-02-21]. https://www. who. int/news/item/01-02-2024-global-cancer-burden-growing-amidst-mounting-need-for-services.

[2] 回文宇,吴锦阳,黄建华,等.机器人辅助颏成形术行截骨操作的精度评价实验研究[J].上海交通大学学报(医学版),2022,42(9):1347-1352.
(Hui Wen-yu, Wu Jin-yang, Huang Jian-hua, et al. Experimental study on the accuracy evaluation of robot-assisted osteotomy of genioplasty[J]. *Journal of Shanghai Jiao Tong University* (*Medical Science*), 2022, 42(9): 1347-1352.)

[3] Chi J N, Miao J, Chen J H, et al. DSTAN: a deformable spatial-temporal attention network with bidirectional sequence feature refinement for speckle noise removal in thyroid ultrasound video[J]. *Journal of Imaging Informatics in Medicine*, 2024, 37(6): 3264-3281.

[4] Zhou T H, Brown M, Snavely N, et al. Unsupervised learning of depth and ego-motion from video[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, 2017: 6612-6619.

(下转第65页)