<table>
<tr><td rowspan="3">项目简况</td><td>项目名称</td><td colspan="6">固定单目视角下（眼科）手术场景医学图像重建研究</td></tr>
<tr><td>项目类别</td><td colspan="6">（ A ）　　A--创新训练项目；B--创业训练项目；C--创业实践项目</td></tr>
<tr><td>申请资助经费</td><td colspan="3">16000 元</td><td>项目起止时间</td><td colspan="2">2024 年 4 月至 2025 年 6 月</td></tr>
<tr><td rowspan="3">项目负责人</td><td>姓　　名</td><td>董炳闻</td><td>性别</td><td>男</td><td>出生年月</td><td colspan="2">2003 年 10 月 28 日</td></tr>
<tr><td>专业年级</td><td colspan="3">大二</td><td>学院（系、部）</td><td colspan="2">计算机系</td></tr>
<tr><td>学分绩点及专业排名</td><td colspan="3">绩点 3.91，排名 12/123</td><td>电　话</td><td colspan="2">13387586473</td></tr>
<tr><td rowspan="3">项目组主要成员</td><td>姓　　名</td><td>性别</td><td>出生年月</td><td>专业年级</td><td>所在学院（系、部）</td><td colspan="2">项目分工</td><td>签　字</td></tr>
<tr><td>丁昊天</td><td>男</td><td>2004 年 7 月 10</td><td>大二</td><td>计算机系</td><td colspan="2">显微镜深度估计</td><td></td></tr>
<tr><td>刘淦</td><td>男</td><td>2004 年 7 月 29</td><td>大二</td><td>计算机系</td><td colspan="2">构建三维重建渲染模型</td><td></td></tr>
<tr><td rowspan="4">指导教师</td><td>姓　　名</td><td colspan="2">性　别</td><td>出生年月</td><td>职　称</td><td>最高学历</td><td>最后学位</td><td>研究方向</td></tr>
<tr><td>胡衍</td><td colspan="2">女</td><td>1986.10</td><td>副研究员</td><td>博士</td><td>博士</td><td>手术辅助，术中导航</td></tr>
<tr><td>刘江</td><td colspan="2">男</td><td>1968.03</td><td>教授</td><td>博士</td><td>博士</td><td>眼科人工智能</td></tr>
<tr><td>电　话</td><td colspan="3">18819070914</td><td>E-mail</td><td colspan="3">huy3@sustech.edu.cn</td></tr>
</table>

一、立项依据（包括项目的意义，国内外研究现状与存在的问题，自身具备的知识条件,自己的特长、兴趣，相关经历，开展研究的前期准备工作等）

（一）项目简介

基于视觉的三维重建技术主要通过使用相关仪器来获取现实中的物体或场景的数据图像，并对仪器获取的图像数据进行分析与加工，构建出符合计算机逻辑表达的数学模型，从而建立真实物体的三维模型的技术。该技术在实际生活中各个领域，如人工智能，自动驾驶，安防监控，运动目标检测等领域都得到了广泛的应用，同时，因为其具有显示清晰，细节丰富，实时性好等优点，也广泛应用于临床医学，尤其是手术室场景中的影像构建，帮助医生们能够摆脱传统二维影像的束缚，更好地进行病灶结构的提取，判断病灶的位置，从而提高手术的精确度，避免因为迷失方向和手眼失调引起术中失误。

然而，术中图像的三维重建与自然图像的三维重建有较大差异。首先，显微手术场景相机固定不动，视角往往受限且晃动频率大，难以从中提取出有效数据。其次，受限于相机的大小，得出的图像特征点三维坐标 xyz 差异也小，匹配上的点较少，现有算法难以提取出有效的信息。除此之外，手术场景中，组织会因为患者呼吸、器械操作等产生移动，而器械则可能因为医生的操作出现快速移动的情况，组织与器械的动态信息不一致的问题进一步为三维重建的精确程度带来了挑战。再考虑到手术场景下对重建速度的高要求，显微手术场景下的三维重建一直是眼科医疗图像处理的一个难点。

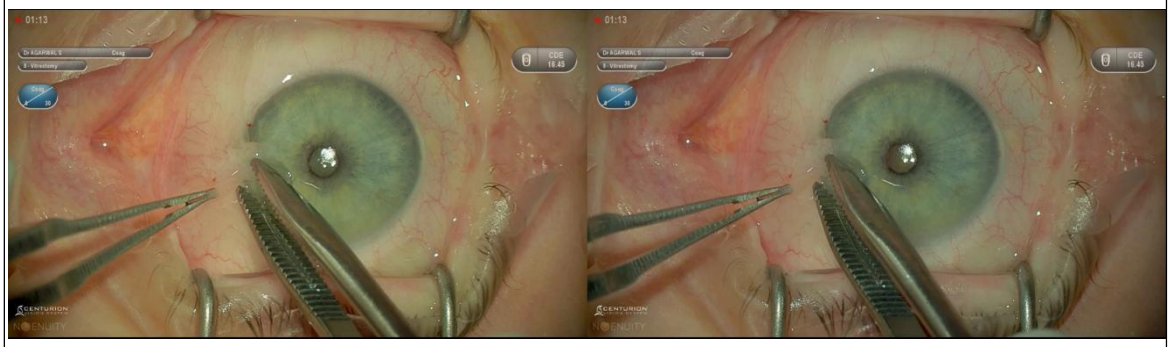


图 1 双目视角下的眼科手术场景

本项目针对视角固定，图像特征点难以提取，组织与器械的动态信息不一致，重建速度慢这四个问题，提出基于事件深度估计的显微手术场景三维快速重建方法。为了更快速的在视角受限且目标运动幅度较大的场景下完成三维重建，本项目引入了基于事件相机的深度估计算法，通过利用事件相机的高时间分辨率和低延迟特性，实现更快速、更准确的深度估计。本项目首先研究此方法在自然场景下的表现情况，进行初步的模型设计与预实验，来验证模型与方法的准确性和重建效率。而后在一些公开的显微手术场景下的数据集上进行实验，逐步对模型结构进行调整和优化。最后将在医院收集的实际数据上进行效果评估。

（二）项目意义

1. 社会意义

眼睛是人体重要的感觉器官，人脑所获得的外界信息中，70%以上来自于视觉。据统计，2019 年全球共有 7.1 亿人患有各种形式的眼病，我国眼病患者约 1.9 亿（27%），居世界之首。眼病不仅给患者带来不同程度的视觉损伤或丧失，导致生活和工作不便，还会加重家庭和社会的负担，是重大的公共卫生问题[1]。

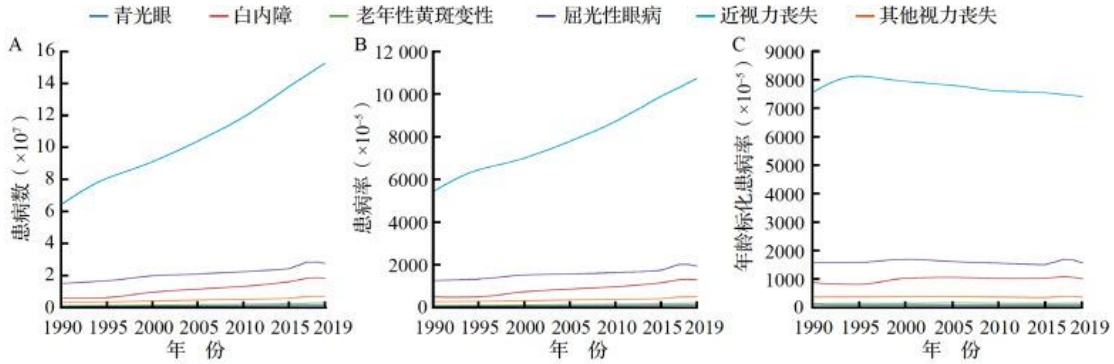


图 2 1990—2019 年中国人群眼病的患病数、患病率和年龄标化患病率

国家眼部疾病临床医学研究中心主任许迅教授也在接受采访时表示，“我国有超过 4000 万的眼底病患者，且医疗资源分布仍不均匀。作为目前国内最主要的致盲性眼病，眼底病的临床现状可以说任务沉重，且对社会生活、经济影响深远。这也是国家为何在今年的‘十四五’眼健康规划中要将眼底病作为新增病种的主要原因。”此外，《“十四五”全国眼健康规划（2021-2025 年）》中还指出，随着经济社会发展及人口老龄化进程加剧，人民群众对眼健康有了更高需求[2]。虽然“十三五”时期，我国眼科医务人员队伍不断完善，眼科医师数量增加至 4.7 万名，但我国眼科优质医疗资源总量相对不足、分布不均衡的问题依然存在，基层眼健康服务能力仍需加强，眼健康工作任务依然艰巨。

为了响应国家号召，考虑到每位眼科医生的临床诊断经验存在差异，诊断的效率与准确率也存在着差异，尤其是一些基层眼科医疗存在力量严重不足，无法提供有效的诊断与筛查，而三甲医院和专科医院眼科服务则“供不应求”，患者无法得到及时的诊疗，容易造成不可挽回的患者眼睛视力损伤和失明。而医学图像三维重建技术可以在手术过程中引导医生对眼球体结构有更精细的了解，从而帮助医生们摆脱传统二维影像的束缚，及时发现并处理可能出现的异常情况。这有助于医生在手术过程中做出及时的调整和优化，确保手术的安全和顺利进行，因此三维重建具有重要的应用价值。

本项目提出以固定单目视角的显微镜下的手术视频为输入，为医生提供高分辨、精细的重建结果。基于事件相机的深度学习技术的引入，一方面可以提高结构的定位精准度，为医生提供更加可靠准确的参照结果，另一方增强了算法的重建速度，使得该算法能够快速得出手术场景的三维重建结果，提高医生执行手术的效率，同时也能够降低手术风险，在最大程度上保障病人的健康。

2. 科学意义

相比于自然场景的三维重建，显微手术场景中医学影像重建的目标通常比较小，光照严重不均衡，且在显微手术场景中，固定的视角和较大的晃动频率都极大影响图像中有效信息的获取。算法在临床上投入使用还需要算法能够实现较高的重建精度和重建速度，因此显微手术场景三维重建的难度更大。现有应用于医学中的三维重建算法大多都针对于脑科和骨科，或者其他内窥镜下的病灶周围模型的建立，更侧重于术前规划、病灶定位、手术导航等方面，无法提供高清晰且精确的三维重建影像结果，并且对于计算资源的需求较高。

而显微手术场景主要用于辅助医生进行精细的手术操作，如修复受损组织、切除病变部位等，对重建出的模型精确度和重建速度的要求极高。在自然场景条件下，通过引入基于时间相机的深度估计算法使得模型能够利用事件相机高时间分辨率和低延迟的特性，在大大提高计算速度的同时能够提供更加精确的模型。但是由于在手术场景中相机视野受限，常常存在体液、血液等液体残留造成遮挡或反光，进一步增加了深度估计的难度和复杂度。目前提出的深度估计算法在面对手术场景时三维重建的效果并不理想，现有方法不适用于该应用场景。

因此，本研究考虑提取场景中的动态细节信息，提出了一种适用于眼科手术显微镜场景的深度估计方法，同时为了提高重建图像的速度与清晰度，提出了一种新的三维染模型，结合输入的前后帧信息，能够快速、清晰的构建三维模型。

（三）国内外研究现状与存在的问题

1. 国内外研究现状

（1）三维重建简介

三维重建指利用二维投影或影像恢复物体三维形状的过程，目前基于深度学习的三维重建算法包含与传统三维重建算法相结合的 DeepVo, BA-Net, CNN-SLAM13[3,4,5]，以及直接利用深度学习进行三维重建的算法。

在传统的三维重建方法中，一定要满足所重建的场景静止、两张图片之间存在重合部分且相机内参已知。已知坐标位置深度 λ 、相机的旋转矩阵 R （3 × 3） 、平移向量 T (3 × 1) 和未处理的点的三维信息$X_i(i = 1,2,\ldots,n)$ (3 × 1) 时， 公式 $\lambda x_i = RX_i + T$ 计算出消除内参的坐标 $x_i(i - 1,2,\ldots,n)$。以$x_i(i - 1,2,\ldots,n)$作为输入，三维重建的最终目的是求出 $R, T, X_1, \ldots, X_n$，并最小化投影误差

$$E(R, T, X_1, \ldots, X_n) = \sum_{i=1}^{n} (||x_1^i||^2 + ||x_2^i - \pi(R, T, X_i)||^2)$$

其中，x_i是两张图片中经过特征点匹配后提取出的特征点，总共能够提取出 n 个共同特征点。在求解三维重建的过程中，首先利用修正后的二维坐标建立对应关系，即先算出关键点，然后再根据所求的关键点进行匹配（可采用 K-最近邻、人工神经网络、布隆过滤器等算法）得到对应关系。接着从提取出的对应关系出发，利用极线约束将姿态和 3D 点云的求解进行解耦，先求姿态（R, T）再求三维点$(X_1, \ldots, X_n)$。

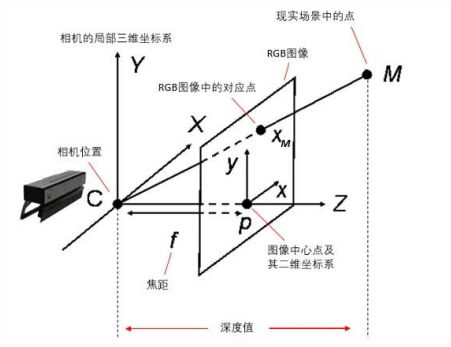


图 3 RGB 图像和深度值对应关系

取有对应关系的一对特征点x_1, x_2，其对应的 3D 点为 X。记坐标位置深度λ_1, λ_2，相机的旋转矩阵 R（3×3），平移向量 T（3×1），根据 RGB 图像和深度值对应关系，可以得到公式 $\lambda_2 x_2 = R(\lambda_1 x_1) + T$。经过变换得到本质约束，即 $x_2^T \hat{T} R x_1 = 0$。其中$\hat{T}R$ 被称为本质矩阵。至此，公式中只剩下姿态（R, T）和位置（x_1, x_2）两个未知量。经过对公式的重组，可以变成 $(x_1 \otimes x_2)^T E = 0$ 形式的公式，其中 E 由本质矩阵扁平化处理后得到的 9 维向量。用 SVD 可以求出 E，即姿态。得到姿态之后，根据线性方程组 $\lambda_2^j x_2^j = \lambda_1^j R x_1^j + \gamma T$，$j = 1,2,\ldots,n$ 可以求解出$(\lambda_1^1, \lambda_1^2, \ldots, \lambda_1^n, \gamma)$，即能够求出 3D 点云。

深度学习为基础的三维建模通常以端到端的方式从图像输入进行三维重建。根据处理的数据形式，深度学习算法可以分成基于深度图、基于体素、基于点云和基于网格的算法。

（2）经典的三维重建方法概述

视角合成是从给定的一组输入图像及其相应的相机姿态中，渲染出场景的新视角。在从新视角生成逼真的图像输出时，需要正确处理复杂的几何结构和材质反射特性。为解决这一问题，已经提出了许多不同的场景表示和渲染方法；然而，传统方法尚无法在大范围的相机视角下实现逼真的质量。目前主流的重建方法是 Neural Radiance Fields(简称 NeRF)[6]和 3D Gaussain Splatting[7]（简称 3DGS）。

NeRF 是一种新的场景表示方法，可以直接优化以重现大量高分辨率的输入视角，并且仍然极具内存效率。

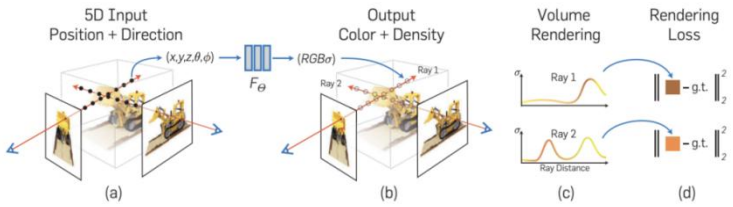


图 4 NeRF 神经网络架构示意图

NeRF 通过神经网络学习三维场景的辐射信息，并利用这些信息来进行高质量图像渲染。它已经在计算机视觉领域取得了许多重要成果。上图是 NeRF 的大致结构，其通过沿着摄像机射线采样 5D 坐标（位置和视角方向）来合成图像，将这些位置输入到多层感知器（MLP）中以产生颜色和体积密度，然后使用体积渲染技术将这些值合成为图像。

另一种经典的三维重建方法 3DGS 是基于点云的数据可视化技术，在三维重建领域中同样具有重要地位。与 NeRF 网络不同，3DGS 基于 Structure from motion(简称 SfM)[8]点云重建三维模型，显示地表达三维信息。3DGS 在渲染的。

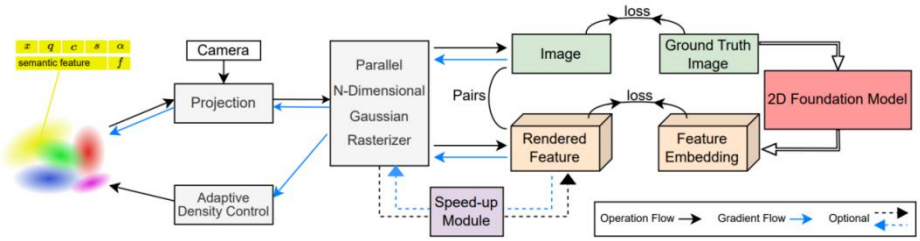


图 5 3DGS 神经网络架构示意图

（3）动态场景的三维重建

相比于静态图片，动态场景指场景中的物体在时间上存在运动或变化的场景。在动态场景中，物体的位置、形状、颜色等属性会随时间而变化，对象的移动也可能导致频繁的遮挡。这些因素使基于动态场景的三维重建的难度远大于基于静态场景的三维重建。早期的基于单摄像机的视觉 SLAM (simultaneous localization and mapping)和基于结构光的方法用于在更少的硬件约束下进行动态场景的捕获和重建。近年来，深度学习在动态三维重建中的应用取得了显著的进展。神经网络能够直接从图像中学习场景的三维结构和动态变化，有优越的性能。

EndoNeRF[9]是医学领域中使用双目相机图像三维重建的方法。通过包括 mask 引导的射线投射、立体深度提示的射线投射和立体深度监督优化等方法，在腹腔手术中在较为有限的数据集上很好地完成了去除手术工具遮挡和场景形变两大任务。

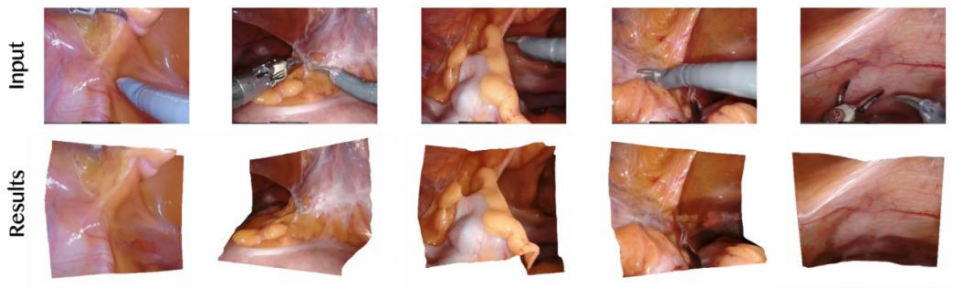


图 6 手术场景下 EndoNeRF 的三维重建示意图

（4）深度估计

深度估计可以分为两类。一类是双目的立体深度估计，这种方法通过视差计后预估图像每个像素点的深度值。另一类是单目深度估计，该方法依赖于一定的几何约束，通过对单张图像进行像素级深度估计。传统上，主要是通过从传感器、立体匹配或 SfM 获取深度数据。这些方法在自然场景中表现良好，甚至有些方法能够预测出非常细腻的深度值，比如 Depth everything[10]。

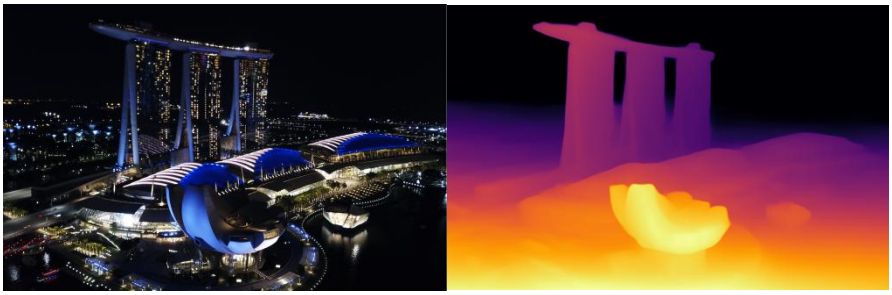


图 7 自然场景实验结果

深度估计在医学场景中应用同样重要。在医学图像的三维重建时，更加准确细腻的场景的深度信息能更好的将整个场景还原。

2. 目前方法存在的问题

（1）眼科手术场景数据集

受限于眼球位置的敏感性，眼球结构的脆弱性，病灶形状的可变性，数据集获取难度较大，公开数据集大部分为如心脏，大脑，胸部，肿瘤等大型器官或病灶的数据集，且大多为手术场景下的内窥镜视角。受限于手术场景的数据质量和完整性，数据标注需要大量专业知识，以及隐私和安全问题，本研究所需要的显微眼科手术场景下可供使用的包含准确真实标签的数据集极少，现有的少数数据集也都有各自的问题，如相机内参数据的缺失等。由于公共数据集的缺乏，对于手术室场景下的三维重建研究发展还未能完全适用于临床环境，因此，研究一种弱依赖医生主观意识的、高精度的、合理自动化水平的医学影像分割与三维重建方法仍有必要且富有挑战。

（2）医学（眼科）显微镜下的相机视角

在当前流行的 NeRF 和 3DGS 的应用中，相机是围绕着场景移动的，可以从多个视角获得充分的信息。然而，在手术场景中，由于手术工具和组织的位置受到操作空间的限制，摄像机的移动范围通常是受限的[9]。因此，内窥镜或显微镜视频往

（3）医学（眼科）显微镜下的深度估计

在医学场景中，当图像聚焦于特定手术部位时，场景中的位置深度差异不大，需要比自然场景更细微的预测，增加深度估计的难度。此外，手术现场常常存在体液、药液等液体残留在机体表面，液体的折射和波动会导致图像扭曲和模糊，进一步增加了深度估计的复杂度[11]。

同时，在显微镜观察中，由于观察位置固定、视线遮挡、色彩平淡以及光线昏暗等因素，画面的深度信息的呈现较为粗糙。因此，将适用于自然环境的深度估计模型应用于医学图像时会面临诸多问题。

图 8 NeRF 训练出来的分辨率较低

（4）医学（眼科）场景的渲染

NeRF 训练出来的场景的分辨率较低。在 NeRF 进行渲染时，需要对场景中的像素进行采样，以确定其颜色和亮度等属性。对于高分辨率的图像或复杂的场景，需要处理大量的数据。这意味着需要大量的计算资源来处理这些数据，从而导致渲染过程变得耗时。虽然现有工作如 InstantNGP[12]很好的解决了渲染的耗时问题，但渲染出来的三维场景仍然十分低质量，这不符合我们对于手术场景中对精细组织结构进行渲染的目标。

（四）前期准备工作

我们阅读了大量的论文，尝试了很多经典的和前沿的研究成果。我们将不同阶段的不同任务分为数据集、mask 方法、深度估计模型和渲染模型四个部分。

1. 数据集

我们在网上查找了大量公开的医学场景数据集资料，从中挑选出了对本研究有所帮助的部分，与实验室所拥有的数据集资料相结合，搭建了本研究所需要的数据集。目前手术场景数据集清单如下表：

数据集名称	数据集类型	格式	时长	帧率	来源
Eyetube	双目眼科手术场景数据	视频	约 10 小时	29.97 帧/秒	eyetube.net
Sim_data	单目眼科虚拟数据集（有深度 gt）	视频	约 10 分钟	29.97 帧/秒	iMED 课题组
Exp_data	双目实验室数据	视频	约 4 分钟	29.97 帧/秒	iMED 课题组

2. 器械分割

因为我们的输入图像中存在手术工具的像素点，这些像素点是无效的，如果这时我们仍在整张图像上均匀随机采样，会出现采样到无效像素点的情况。mask（掩码、掩膜）是深度学习中的常见操作。简单而言，其相当于在原始张量上盖上一层掩膜，从而屏蔽或选择一些特定元素。需要生成 mask 掩膜来表示手术工具在视频中的位置。我们采用了常用的图像分割方法：阈值分割和边缘检测，对手术场景下的图片尝试进行了 mask 掩膜的生成。如图所示，我们已经用生成了以下 mask 掩膜，白色区域为手术工具。

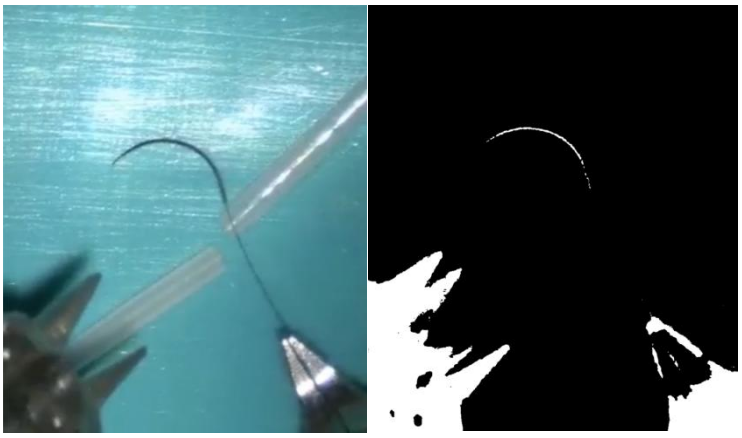


图 9 mask 掩膜实验效果展示

3. 深度估计

我们尝试用传统机器学习和深度学习不同的深度预测方法对手术场景的图片进行了深度预测。传统机器学习方法中最有代表性的是 SfM 方法，在多视角的场景中

无法被准确地匹配和重建。SfM 形成的不完整重建在视觉上不能构建出场景的轮廓，生成的模型没有规律，表现很差。在单目或双目的约束下，深度学习的深度估计模型表现相对较好。基于神经网络的研究 Depth everything 的实验结果能够大致反映出眼底手术场景的轮廓，对手术器械的深度估计也达到了较好的水平。但是该模型也不能很好的反映出眼球底部细腻的的深度纹理。如图 10 所示，眼底的深度信息呈现深紫色，这说明眼底深度信息仍然模糊，眼底的位置、形态仍然估计的不好。因此，三维重建需要更好的深度信息图。

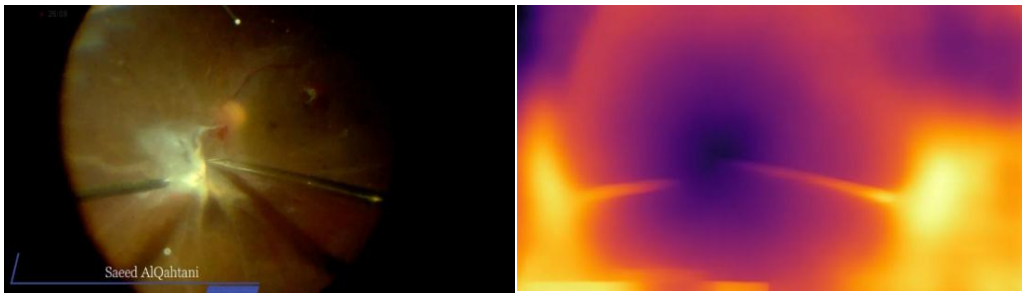


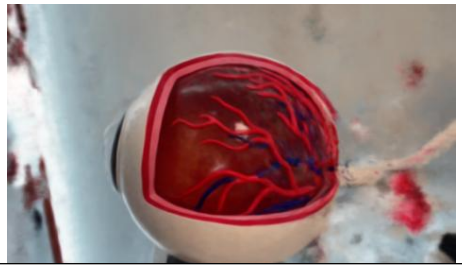
图 10 深度估计效果图

4. 渲染模型

我们测试了目前主流的几个三维重建模型，对于一眼球模型进行三维重建。

我们尝试了 Nerf， 3DGS 和 4DGS(4D Gaussian Splatting)三种渲染方法[6][7][14]，前面两种模型需要多角度的图片输入 4DGS 则需要输入图片。在多视角的输入下，3DGS 比 Nerf 效果更好。而 4DGS 的效果优于 3DGS。

我们初步用 NeRF 对眼球模型进行重建。我们首先环绕眼球模型进行环绕拍摄，进而将拍摄获得的照片用 Colmap[8,14]生成相机位参信息，并进行稀疏三维重建。我们将生成的相机位参和稀疏重建信息输入神经网络。下图是 NeRF 渲染两分钟后的结果，已取得不错的效果，有明显的三维感。



3DGS 效果图如下，相比 NeRF，3DGS 对场景的形状、纹理细节有更好的捕捉。同时，3DGS 的渲染速度比 NeRF 提升了很多。不足的是，3DGS 生成的三维模型仍然存在模糊的杂质。

图 12 3DGS 对眼球模型的三维建模

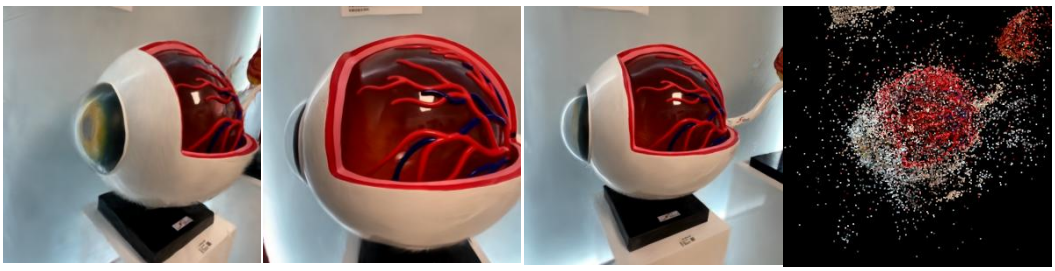


图 13 4DGS 对眼球模型的三维建模

如图 13 所示，4DGS 建模的效果图如左三所示，右一是经过算法稀疏重建后的点云图。可见，相比于 3DGS，4DGS 得到的结果点云更加密集，渲染质量更高。但是，4DGS 在医学场景中任然存在一些问题。首先，4DGS 是通过多视角的连续图像输入得到的结果，然而在医学场景中，只有单目输入。其次，4DGS 的渲染速度不快，对于可能有半个小时甚至几个小时的长时间手术视频，如果使用 4DGS 进行渲染，则渲染模型的时间成本会非常高。同时，在手术场景中，光线的问题也会影响渲染结果。因此，本项目将尝试通过单目视频建立点云，研究在医学场景中对渲染过程质量和速度的优化方法。

（五）自身具备的知识条件

我们三人均就读于南方科技大学计算机科学与工程系，有良好的计算机基础，熟悉 Java，C++，Python 等多种语言以三维建模的诸多基础算法。

了与真实手术场景较为接近，但由于场景光线较暗，深度差异不大，场景中有液体等因素导致现有模型表现不好的数据集进行三维重建和模型训练工作。

我们在日常的学习与思考中也产生了一个较为清晰的思路。在我们研究的过程中，课题组可以提供优秀的数据资源和计算资源进行辅助。

二、项目研究内容

（一）具体研究内容和技术考核指标

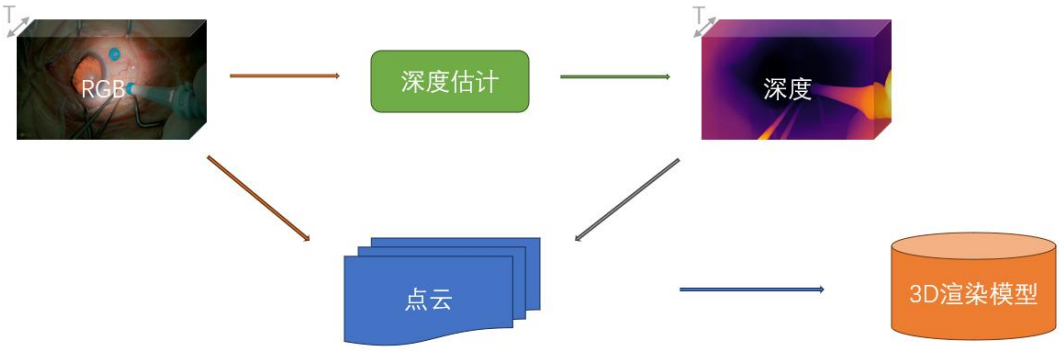


图 14 项目思路示意图

本项目研究如何通过输入单目视频进行医学场景的三维重建。图 14 中，T 表示采集的视频帧数，通过深度信息构造的点云模型与前后一共 T 帧的点云模型结合作为 3D 渲染模型的输入（动态信息包含在前后时间不同帧的关系中），得到最终结果。具体的研究内容包括:

1. 眼科手术视频数据集的构建

此项目将在现有的眼科医学的医学场景数据集和眼科虚拟手术集的基础上，整理成三维重建所需要的数据集，并在此数据集上进行深度标定。此外，项目还将人工标定使用于拍摄过程中的显微镜的内参。在数据集上将评估预处理算法的噪声抑制效果和图像校正效果作为数据集准确性指标。

2. 基于事件信息的单目手术视频的深度估计

项目中将研究针对医学手术场景更有效的深度估计模型，对输入的单目手术视频的每帧图像进行更精确的深度预测。在手术场景中，显微镜场景固定，基于事件相机（event-based）的方法适用于这种状态下的深度估计。事件相机是异步传感器，它基于场景动态采样光，而不基于时间[15]。具体来说，在事件相机的处理中，当该像素处的亮度变化超过设定阈值时，就会生成、输出事件数据。设定亮度为 I，定义亮度为实际亮度的对数值，$L = \log(I)$，那么，在t_k时刻$x_k = \{x_k, y_k\}$像素处的亮度增益为

$$\Delta L(x_k, t_k) = L(x_k, t_k) - L(x_k, t_k - \Delta t_k)$$

其中p_k表示为事件极性。基于 RGB 影像中前后帧的亮度差异进行阈值处理即可将 RGB 视频转化为基于事件的视频数据。本项目计划提出基于事件相机的深度估计的模型。其输入是模拟事件相机的数据，输出是每帧的深度估计图像。这种方法能够克服低亮度、运动模糊等问题，适用于手术场景。

最后，将深度图像帧、RGB 图像帧和相机的内外参信息综合，共同推断出场景的三维点云模型。深度估计的考核指标为：在有标签数据中深度估计的考核指标有绝对误差(AbsRel)、均方根误差(MSE)、深度准确率(Acc)和与真实标签的一致性等指标考核深度估计的效果，对于无标签的数据，通过观察深度图像轮廓、平滑度来判断深度估计结果。

相机内外参的计算指标有重影误差指标、畸变矫正效果、相机姿态稳定性等。对于生成的点云，我们用点云密度、均匀性、准确性、平滑度、完整性以及可视化效果判断生成点云质量。

3. 快速三维重建算法研究和优化

项目旨在开发一种新颖的三维重建算法，受到 4DGS 的启发，我们能够利用时序点云数据进行四维（3D+时间）重建，以获得更好的效果。通过将时间维度整合到现有的三维模型中，计划实现动态的重建效果。特别地，该算法将针对眼科医学领域的特定需求进行优化，考虑到医学图像的独特数据特性和实际应用场景的要求。本项目还将优化三维重建方法，以提高重建的速度和精度。技术考核指标有三维重建算法的重建精度和效率，以及算法在医学应用中的实际可用性、稳定性。

在实现动态三维重建后，首先我们将对于环境光线不均匀的问题进行优化。在手术场景中，由于显微镜的结构限制和手术场景的特殊性，在手术场景中获取的图像通常会存在光线不均匀的问题。光线不均匀会导致三维模型的渲染过程生中会有许多“白雾”产生，如图 11 所示。Duckworth[16]提出了一种对光线问题处理的方法。通过对 3DGS 密度的自适应控制实现。根据其研究，经过 30K 次迭代（约 35 分钟），背景伪影已经得到了显著减少。

（二）拟解决的关键问题

1. 显微镜下的医学场景视角固定，项目首先需要获取基于事件的视频影像。然后将基于事件的深度估计模型在医学场景下做改进和适配，克服显微镜下医学影像数据低光照和对比度高的特点，进行细致的显微镜场景深度估计。

2. 由深度信息和 RGB 信息提取显微镜场景的三维信息是本项目的另一个需要克服的难点。相机输入视角单一，通过单目图像提取场景中的三维信息具有挑战。同时显微镜下场景狭小，眼底的深度差异很小，也会导致提取信息不准确。项目需要在提取细致准确的三维信息时三维克服这些问题。

3. 将时序信息融入传统的三维重建模型，使三维重建在医学场景下的效果更加清晰。项目将仔细研究如何将时序信息融入三维重建模型，这种方法应用在医学显微镜手术场景是创新的。所以如何将医学显微镜手术场景的时序信息有效的利用以构建出四维的重建模型将会是本项目中需要解决的问题。

4. 加速模型渲染速度，减少三维模型建立的时间成本。现有的三维渲染速度普遍很慢，项目中计划优化三维重建的渲染模型，让渲染的速度得到提升。这是项目中拟解决的关键问题。

（三）项目可行性分析

从原理上来说，三维重建技术已经在多个领域展现出其强大的潜力和实用性。医学领域，特别是眼科手术场景下的挑战包括但不限于小范围的操作空间、细微的手术细节和特殊的光照条件。尽管如此，基于现有的研究和技术，通过对深度学习模型和算法的优化，可以实现对单目视频中眼科手术场景的三维重建。此外，通过集成时间维度信息，可以进一步提升重建模型的动态表现和精度。因此，从理论和技术原理上讲，项目的核心目标是可行的。

从实验条件上来说，本项目数据采集所需的手术显微镜（上海轶德公司 SM2000J 实验手术显微镜）和精密移动平台（大恒光电 GCM-T 系列精密平移台，大恒光电 GCM-V 系列精密侧升降台）都已经具备。手术显微镜能够提供清晰的放大图像，帮助医生进行精确的手术操作。精密移动平台可以用于精确地移动样本或仪器，以进行观察和数据采集。实验场地在 iMED 深圳团队南方科技大学实验室。

从项目组员上来说，项目组员有三人都有过项目经历，对本项目有浓厚的兴趣且都在 iMED 深圳团队。 iMED 深圳团队是一个专注眼科人工智能的国际化知名团队，指导老师胡衍博士、刘江老师也在手术辅助、术中导航等研究方向深耕多年，能够给团队成员提供有力的指导。

三、项目实施方案（包括 1.项目人员组成及分工；2. 项目研究进度安排：包括查阅资料、选题、自主设计项目研究方案、开题报告、实验研究、数据统计、处理与分析、研制开发、填写结题表、撰写研究论文和总结报告、参加结题答辩和成果推广；3.配套条件要求：包括仪器设备、场地、材料、资料、实验课时及指导教师要求等）

（一）项目人员组成及分工

董炳闻，项目负责人。主要负责文献查找、论文撰写、模型实验、模型调整、数据整理以及模型的整合。

刘淦，项目参与人。主要负责文献查找、论文撰写、渲染模型构建以及数据整理。

丁昊天，项目参与人。主要负责文献查找、论文撰写、数据集获取、深度估计模型构建。

（二）研究进度及安排

2024.04　进一步进行相关文献调研，数据集构建

2024.05-2024.06 制定研究方案，对后续实验进行规划

2024.07-2024.01 完成模型的初步实验，对模型进行调整优化，专利申请

2024.01-2025.03 进行对比实验，进一步调整模型

2025.03-2025.06 撰写结题报告与论文

（三）配套条件要求

设备：上海轶德公司 SM2000J 实验手术显微镜，大恒光电 GCM-T 系列精密平移台，大恒光电 GCM-V 系列精密侧升降台，四张 NVIDIA RTX 2080Ti 显卡用于模型训练，一台移动医疗设备用于测试算法的实际性能。

场地：iMED 深圳团队南方科技大学实验室

材料：假眼模型，猪眼模型，虚拟眼球模型，真实手术场景数据

四、项目经费预算（包括大概支出科目、金额、计算根据及理由）

序号	支出科目	预算金额	计算根据及理由
1	调研	500 元	本项目相关文献的购买和下载
2	资料、复印	500 元	硒鼓、墨盒、打印、复印费等
3	实验材料	2000 元	购买硬盘、眼科显微手术器械等
4	文献/知识产权事务费	6000 元	计划申请两项专利
5	上机机时费	1000 元	购买云计算服务器
6	发表论文	6000 元	计划投稿一篇国际/国内会议论文
合计		16000 元	

五、预期成果（研究论文、专著、调研报告、设计方案、专利、软件、产品、成果鉴定等）

预期至少发表期刊或会议论文一篇，至少申请一项专利，打造一套方便使用的医学（眼科）三维建模平台。

六、本项目的创新之处

1. 提出一种基于事件相机的深度估计方法

一般的深度估计方法无法精确捕捉医学场景中细微的深度差异，没有办法很好的处理体液、药液以及其他液体对于深度图像的影响。由视频输入的深度估计模型通常是相机和物体同时运动，而再显微镜视角下，相机位置固定，手术场景的运动幅度和规律也不同于自然场景。所以这些深度估计方法不适用于本项目针对的场景。因此我们基于事件相机的输入，提出了一种适用于眼科手术显微镜场景的深度估计方法，具有创新型。

2. 提出一种新颖的结合时序信息的三维渲染方法

常见的三维渲染方法基于多视角输入提取的点云模型，在显微镜视角下，仅有单目视频的输入。这些三维渲染的方法针对的都是静态的三维场景，没有很好的利用前后帧连续的空间信息。我们提出了一种新的三维渲染模型，结合输入的前后帧信息，能够快速、清晰的构建三维模型。

3. 提出一种给三维渲染模型加速的优化方法

目前流行的 NeRF、3DGS 方法的速度很慢，其渲染时间与输入的图片数量成正比。对于可能长达几个小时的医学视频输入，渲染的时间成本巨大。项目计划提出一种能够加快三维渲染速度的方法，能够快速的对长时间的医学视频输入处理，构建其时序的三维重建模型。

参考文献：

[1] Chen, B., Lou, L., & Ye, J. (2021). Eye diseases burden in China in the past 30 years. Zhejiang da xue xue bao. Yi xue ban = Journal of Zhejiang University. Medical sciences, 50(4), 420–428. https://doi.org/10.3724/zdxbyxb-2021-0246

[2] Zhu, Y. (2022, January 4). Notice of the National Health Commission on Printing and Issuing the “14th Five-Year” National Eye Health Plan (2021-2025). 国家卫生健康委关于印发"十四五"全国眼健康规划（2021-2025 年）的通知_国务院部门文件_中国政府网. https://www.gov.cn/zhengce/zhengceku/2022-01/17/content_5668951.htm

[3] Wang, S., Clark, R., Wen, H., & Trigoni, N. (2017, September 25). Deepvo: Towards end-to-end visual odometry with deep recurrent convolutional neural networks. arXiv.org. https://arxiv.org/abs/1709.08429

[4] Tang, C., & Tan, P. (2019, August 25). Ba-net: Dense Bundle Adjustment Network. arXiv.org. https://arxiv.org/abs/1806.04807

[5] Tateno, K., Tombari, F., Laina, I., & Navab, N. (2017, April 11). CNN-Slam: Real-time dense monocular slam with learned depth prediction. arXiv.org. https://arxiv.org/abs/1704.03489

[6] Wang, Z., Wu, S., Xie, W., Chen, M., & Prisacariu, V. A. (2022, April 6). Nerf--: Neural radiance fields without known camera parameters. arXiv.org. https://arxiv.org/abs/2102.07064

[7] Kerbl, B., Kopanas, G., Leimkühler, T., & Drettakis, G. (2023). 3d gaussian splatting for real-time radiance field rendering. ACM Transactions on Graphics, 42(4), 1-14.

[8] Schonberger, J. L., & Frahm, J.-M. (2016). Structure-from-motion revisited. 2016

IEEE Conference on Computer Vision and Pattern Recognition (CVPR). https://doi.org/10.1109/cvpr.2016.445

[9] Wang, Y., Long, Y., Fan, S. H., & Dou, Q. (2022, June 30). Neural rendering for stereo 3D reconstruction of deformable tissues in robotic surgery. arXiv.org. https://arxiv.org/abs/2206.15255

[10] Yang, L., Kang, B., Huang, Z., Xu, X., Feng, J., & Zhao, H. (2024, April 7). Depth anything: Unleashing the power of large-scale unlabeled data. arXiv.org. https://arxiv.org/abs/2401.10891

[11]Niemeyer, M., Manhardt, F., Rakotosaona, M. J., Oechsle, M., Duckworth, D., Gosula, R., ... & Tombari, F. (2024). RadSplat: Radiance Field-Informed Gaussian Splatting for Robust Real-Time Rendering with 900+ FPS. arXiv preprint arXiv:2403.13806.

[12] Müller, T., Evans, A., Schied, C., & Keller, A. (2022, May 4). Instant neural graphics primitives with a multiresolution hash encoding. arXiv.org. https://arxiv.org/abs/2201.05989

[13] Wu, G., Yi, T., Fang, J., Xie, L., Zhang, X., Wei, W., Liu, W., Tian, Q., & Wang, X. (2023, December 7). 4d gaussian splatting for real-time dynamic scene rendering. arXiv.org. https://arxiv.org/abs/2310.08528

[14] Schönberger, J. L., Zheng, E., Frahm, J.-M., & Pollefeys, M. (2016). Pixelwise view selection for unstructured multi-view stereo. Computer Vision – ECCV 2016, 501 – 518. https://doi.org/10.1007/978-3-319-46487-9_31

[15] Gallego, G., Delbruck, T., Orchard, G., Bartolozzi, C., Taba, B., Censi, A., Leutenegger, S., Davison, A. J., Conradt, J., Daniilidis, K., & Scaramuzza, D. (2022). Event-based Vision: A Survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(1), 154 – 180. https://doi.org/10.1109/tpami.2020.3008413

[16] Duckworth, D., Hedman, P., Reiser, C., Zhizhin, P., Thibert, J.-F., Lučić, M., Szeliski, R., & Barron, J. T. (2024, February 6). SMERF: Streamable memory efficient radiance fields for real-time large-scene exploration. arXiv.org. https://arxiv.org/abs/2312.07541