小 型 微 型 计 算 机 系 统
Journal of Chinese Computer Systems
DOI:10.20009/j.cnki.21-1106/TP.2025-0107
2026 年 5 月 第 5 期
Vol.47 No.5 2026

嵌入双重注意力机制的自监督单目内窥镜深度估计

张连武,李 胜

(浙江工业大学 信息工程学院,杭州 310023)
E-mail:3463126754@qq.com

摘 要: 在内窥镜场景下组织表面纹理稀疏且视野受限,显著增加了深度估计难度.传统方法易受噪声、纹理缺失及光照变化干扰,导致结果稳定性不足.为提高内窥镜图像深度估计的准确性,提出了一种嵌入双重注意力机制的自监督单目内窥镜深度估计网络架构.该网络采用编码器-解码器结构,为了提高模型的准确性,本文在网络架构中集成了双重注意力机制,具体包括通道注意力和空间注意力模块,用以在通道和空间维度上提取远距离的上下文信息.同时引入光度重投影误差和结构相似性和边缘感知平滑作为损失函数,以适应内窥镜图像的特殊属性.最后在 Endoslam 公共数据集进行测试,结果表明本文所提方法能够有效提高内窥镜图像深度估计的准确性.

关 键 词: 内窥镜图像;单目深度估计;通道注意力;空间注意力;自监督学习
中图分类号: TP391 **文献标识码:** A **文 章 编 号:** 1000-1220(2026)05-1212-07

Self Supervised Monocular Endoscope Depth Estimation with Embedded Dual Attention Mechanism

ZHANG Lianwu, LI Sheng
(College of Information Engineering, Zhejiang University of Technology, Hangzhou 310023, China)

Abstract: The sparse texture and restricted field of view of the tissue surface in endoscopic scenes significantly increases the difficulty of depth estimation. Conventional methods are susceptible to interference from noise, missing texture and illumination variations, resulting in insufficient stability of the results. To improve the accuracy of endoscopic image depth estimation, a self-supervised monocular endoscopic depth estimation network architecture embedded with a dual attention mechanism is proposed. The network adopts an encoder-decoder structure, and in order to improve the accuracy of the model, this paper integrates a dual-attention mechanism in the network architecture, which specifically includes channel attention and spatial attention modules for extracting contextual information at a distance in both channel and spatial dimensions. Meanwhile, photometric reprojection error and structural similarity and edge-aware smoothing are introduced as loss functions to accommodate the special properties of endoscopic images. Finally, it is tested on Endoslam public dataset, and the results show that the method proposed in this paper can effectively improve the accuracy of depth estimation of endoscopic images.

Keywords: endoscopic images; monocular depth estimation; channel attention mechanism; spatial attention mechanism; self-supervised learning

0 引 言

近几十年来,随着社会经济发展与健康意识增强,医疗技术革新推动内窥镜成为现代诊疗的重要工具.该设备通过微型创口对人体腔道进行可视化探查,极大降低了传统检查带来的组织损伤,尤其在微创手术中,医生借助实时传输的内窥镜影像精准定位术区,实现创伤最小化的器械操作.尽管如此,内窥镜技术仍面临三维空间感知受限、术野狭窄及器械运动自由度不足等固有缺陷.虽然部分问题可通过强化医师培训得到缓解,但影像深度信息的量化解析始终难以依靠主观经验实现.为此,学界近年聚焦于内窥镜深度重建算法开发[1,2],其技术成果已有效集成于智能手术导航平台[3].

与自然图像相比,内窥镜中观察到的纹理特征更稀疏,更不均匀,使得网络难以从中获得可靠的信息[4].内窥镜深度估计方法根据成像系统配置差异主要划分为双目与单目两大技术路线[5],其临床价值在于通过三维场景重构提升术野认知精度,并为外科手术提供关键导航支持.

目前基于双目视觉的深度估计已经有许多研究.双目立体匹配作为深度感知的核心技术手段,其原理基于对左右双摄同步采集图像的差异分析来推算物体深度信息.经典立体匹配框架遵循 4 阶段基准架构:代价计算、聚合操作、视差推断与优化模块.虽然这类方法在标准场景中保持可行性,然而遭遇高反射表面、弱纹理区域或重复结构条件时呈现精度退化与鲁棒性劣化.这种技术瓶颈推动了基于深度学习的新型视差图回归方法的快速发展.Mayer 团队首创编解码架构 DispNet[6],通过端到端学习实现双目图像对视差图的直接映射,首次将传统四阶段立体匹配流程压缩至单网络框架.Pang 等人在 DispNet 网络基础上提出 CRL 网络结构[7].Kendall 等

收稿日期:2025-03-13 收修改稿日期:2025-04-23 基金项目:国家基金重点项目(62233016)资助. 作者简介:**张连武**,男,1999 年生,硕士研究生,研究方向为深度估计;**李 胜**,男,1984 年生,博士,副教授,研究方向为图像处理、压缩感知、推荐系统等.

人开发了一种用于双目视差估计的新型端到端框架(GC-Net)[8]. 这些方法通过深层架构堆叠与复杂特征融合机制实现视差预测精度增益,但伴随模型复杂度的指数级增长,推理效率呈现显著衰减. 并且双目内窥镜复杂度高、体积大,在人体内使用不便.

幸运的是,近年来的单目深度估计技术已经显示出与传统立体深度估计方法相当的性能. 因此,单目深度估计方法已成为内窥镜检查的主流. 当前单目内窥镜深度估计技术体系主要涵盖两大研究方向:传统计算机视觉方法与深度学习框架. 传统方法以同步定位与建图(SLAM)[9]和运动恢复结构(SFM)[10]为代表,通过分析单目视频序列或多视角图像数据,运用特征点匹配与三角测量原理联合推算场景深度及摄像机运动参数. 这类技术因无需外置传感器且实施成本较低,已成功应用于胃部检测、腹腔手术等多种临床场景的三维重建与运动追踪[11],但在组织表面特征稀疏区域易产生非均匀点云分布. 针对该缺陷,研究团队通过融合激光扫描仪与视觉SLAM 技术实现了组织表面稠密重建[12],但是该方法重建的结果存在细节丢失的问题,无法反应组织表面特性.

深度学习方案分有监督和无监督两类. 有监督网络在常规场景深度预测中表现优异[13,14],但在人体内应用时因无法部署 LiDAR 等深度传感器而面临监督信号缺失的根本问题. 文献[3]提出基于 CT 三维重建与数字孪生技术生成合成深度图,但该方法不仅需要术前 CT 数据且会弱化组织纹理细节,对低纹理的消化系统组织尤为不利. 自监督学习作为无监督子类,通过预训练任务从未标注数据自主生成监督信号,获得可迁移的通用特征(如利用无标签医学影像学习),其优势在于降低数据成本并提升工程可行性. 文献[15]开发了基于单目内窥镜视频的自监督框架,通过多视角立体视觉生成稀疏监督信号实现密集深度预测,无需人工标注或 CT 数据. 然而,由于人体组织形态与光学特性的高度异质性,该算法在临床影像中的泛化能力仍面临显著挑战. 文献[16]提出轻量级方案 EndoDepthL,融合多尺度扩张卷积与多通道注意力机制优化 CNN 架构,通过局部与全局特征协同提取提升深度估计精度. 文献[17]提出了一种多尺度残差融合方法来估计单目内窥镜图像的深度,通过利用图像频域分量空间变换来解决相干照明问题,从而增强场景光源的稳定性,但该方法难以提供空间一致性,并且通常没有足够的上下文意识. 文献[18]提出了一种融合语义分割的自监督内窥镜深度估计框架,基于诊断文本和内窥镜图像之间的对应关系,从而间接提高预测深度的准确性. 但该方法未给出更全面的评估,且所获得的深度图仍存在细节缺缺失的问题. 研究者们提出了多种创新方法来提升内窥镜影像深度感知精度,然而现有解决方案在复杂腔内环境下的解剖结构重建中仍面临核心挑战,特别是在处理组织形变与光照变异场景时,其空间坐标恢复的鲁棒性尚存明显技术瓶颈. 基于自监督的内窥镜图像深度估计是一种有吸引力的替代方法,但它也提出了一系列挑战,除了估计深度,该模型还需要估计训练期间时间图像对之间的自我运动. 结果表明,基于自监督方法的单目深度估计效果明显不如全监督方法,尤其是在目标边缘和弱纹理区域.

本文提出了一种用于内窥镜检查的具有双重注意力机制的自监督单目深度估计方法,以使模型更有效地利用全局信息,从而提高模型的估计精度. 本文还引入了边缘感知平滑误差、多尺度结构相似性和光度重投影误差作为损失函数以提高性能. 总的来说本文有以下创新点:

1)提出了一种自监督神经网络框架,该框架利用并行预测分支获得的深度信息和位姿信息重构图像,并将重构图像作为自监督信号指导网络模型训练. 为了提高自监督模型的准确性,本文在网络结构中嵌入了双重力注意机制,即通道注意力和空间注意力模块分别在通道和空间维度上捕获远程情境信息.

2) 本文还引入了边缘感知平滑误差、多尺度结构相似性和光度重投影误差作为损失函数,以适应内窥镜图像的特殊属性. 例如光照变化和有限纹理,有利于保留高频信息,并能保持亮度和颜色的不变性,从而提高了模型表示图像细节的能力.

1　单目深度估计模型

1.1　总体深度估计模型框架

内窥镜图像的深度估计对于构建内窥镜环境的精确 3D 表面模型至关重要,这可以为医生提供更好的手术部位的理解. 在本节中,本文提出了一种利用双重注意力机制融合结构优化的单目内窥镜自监督图像深度估计方法. 与有监督方法不同的是,本文的自监督架构使用两个网络来进行学习:第 1 个是具有双重注意力机制的深度估计网络,用于学习深度,第 2 个是具有双重注意力机制的位姿估计网络,用于学习预测两帧之间的相对相机运动. 图 1 显示了整个网络架构,深度估

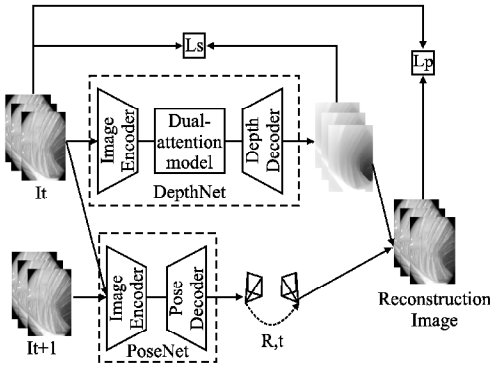

图 1　总体网络结构图

Fig. 1　Overall network architecture diagram

计网络将单个 RGB 图像作为输入,并输出输入图像的估计深度图. 位姿估计网络的输入是两帧图像,输出是 6 自由度位姿变换(相机坐标系和世界坐标系之间的变换是欧几里得变换,其包括旋转和平移. 估计的位姿和深度与相机的内部参数相结合,并被重新投影回平面上以生成重构图像. 以此方式,可以基于重构图像与原始图像之间的相似性来隐式地约束深度网络和姿势网络. 同时利用光度重投影误差、结构相似性和边缘感知平滑误差对位姿估计网络和深度估计网络进行训练.

1.2　自监督单目内窥镜深度估计算法原理

自监督学习深度估计方法不需要依赖真实深度图像作为监督信号,而是采用没有真实值的图像对或视频序列训练网

络,将深度估计看作是一项图像重建任务,并通过最小化重建图像的损失函数. 自监督深度估计框架采用单目视频流进行模型训练,其核心在于在相邻帧之间的投影上构建了几何约束:网络架构以时序连续的单目影像序列为输入源,通过相邻视角间的像素级投影变换生成自监督信号,其中视图合成任务的几何误差被转化为网络优化的目标函数. 由网络模型输出的深度图和姿态变换被用作图像重建的中间变量. 如果给定源帧 I_s 和目标帧 I_t,从网络模型中可以得到估计的深度图 D_t 和估计的从 $I_t \sim I_s$ 的坐标变换矩阵 $T_{t\to s}$,则 I_t 和 I_s 之间的像素投影关系可以由公式(1)表示:

$$p_s = KT_{t\to s}D_tK^{-1}p_t \tag{1}$$

其中 p_s 和 p_t 分别是源帧 I_s 和目标帧 I_t 的像素坐标,K 摄像机固有矩阵,$T_{t\to s}$为从目标帧到源帧的相机运动矩阵,D_t 表示估计的深度图.

$$K = \begin{matrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{matrix} \tag{2}$$

其中,f_x、f_y 是焦距,c_x、c_y 表示主点.

对于单目训练,本文使用两个连续帧分别作为源图像和目标图像. 因此,可以通过网络模型来学习每个像素的深度和帧之间的姿态变换,然后可以根据公式(1)来计算投影关系,利用双线性采样可以实现对源帧的重构. 将目标帧与重建图像的对比度误差作为损失的一部分来训练网络,使得模型能够进行自监督训练,从而在不需要实际摄像机运动的姿态标注和深度的地面真实度的情况下完成深度估计任务.

1.3 深度网络的结构

本文设计的深度网络基于 U-Net[20,21]架构,包括跳跃连接、通道注意力模块和空间注意力模块,能够融合局部深度特征并捕获全局上下文信息. 深度估计网络采用编码器-解码器架构框架,其结构示意如图 2 所示. 为提高模型的训练鲁棒性,编码器选用改进型 ResNet-18[22],该结构通过残差连接机制实现跨层信息传递,在缓解深层网络梯度消失问题的同时,兼具参数量少和推理速度快的优势. 具体而言,编码器以单目 RGB 图像作为输入,首先通过卷积层与批归一化(BN)层进行特征处理:BN 层对特征数据进行标准化,有效抑制梯度异常并加速训练收敛;随后经 ReLU 非线性激活后接入最大池化层,最大池化层对提取特征压缩,简化网络复杂度. 为了提取内窥镜图像的深度特征和增强特征之间的相关性,本文在编码器中嵌入了空间注意力模块 PA 和通道注意力模块 CA,用于探索全局图像区域,以估计这些区域的相似深度. 本文的研究采用在 ImageNet 上预训练的权重进行编码器参数初始化,实验结果显示,相较于从零开始训练的基准方法,该初始化策略提高了模型的准确率. 随后特征图进入 Layer1 模块,Layer1 的结构如图 2 所示,网络层级架构中 Layer1 由双残差模块构成,后续 Layer2 ~ Layer4 均采用相同拓扑结构. 并通过逐级下采样操作,特征图的空间分辨率呈等比缩减,最终由编码器完成多尺度特征的提取过程. 为了进一步增强图像特征之间的相关性,本文在编解码器架构的过渡阶段嵌入双重注意力机制,通过同步捕获通道与空间维度的交互关系,有效强化特征表达的关联性. 解码器通过多级特征融合与解析实现图像重建,其网络包含 4 个同构上卷积模块(Upconv),上卷积模块(Upconv)的结构详见图 2. 各个模块采用双路特征融合机制:将前级输出特征与编码器对应尺度的特征图通过跳跃连接进行通道拼接,将相同尺寸的特征图融合后,进行卷积操作、上采样操作,最终实现与输入图像等分辨率的预测输出. 该解码器在特征重构阶段采用 ELU 激活函数[24],为了将输出限制在合理的范围内,在每个 Upconv 的输出端使用 Sigmoid 函数行数值规约,对输出执行线性变换:

$$D = a + (B - a)\mu \tag{3}$$

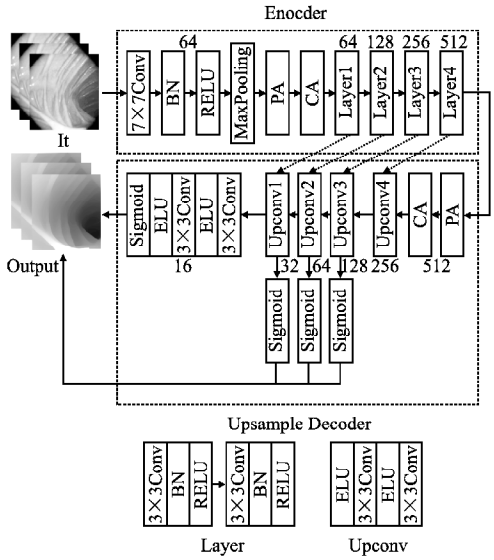

图 2 深度网络结构图

Fig. 2 Deep network architecture diagram

其中 D 是模型输出的深度图,μ 是 Sigmoid 激活函数的输出,a 和 B 分别是最小和最大深度值. 使用公式(3)将最后一个 Sigmoid 函数输出结果转换为深度,选择和将约束在 0.1 ~ 100 个单位之间. 并且创新性地引入反射填充策略替代传统零填充,有效抑制特征图边界伪影,提升重建特征的边缘锐度与结构完整性.

1.4 位姿估计网络的结构

该位姿估计模型基于 CNN 编码器-解码器框架构建(如图 3 所示). 由于单帧 RGB 图像无法提供场景的三维几何表征,所以本研究通过单目视频序列中连续帧间的时序运动关

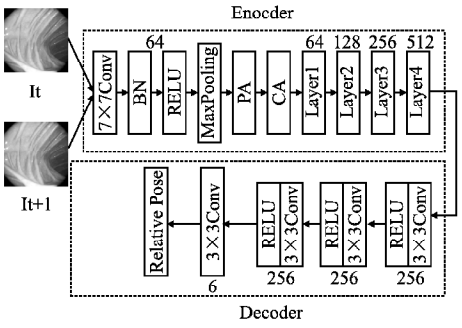

图 3 位姿网络结构图

Fig. 3 Position network structure diagram

联建立多视角约束,进而推断相机的位姿变化参数. 位姿网络以时序相邻的 RGB 图像对(上一帧与当前帧)作为输入,是

一对彩色图像,所以位姿网络接收 6 通道作为输入. 位姿网络与深度网络在架构流程上具有相似性,其网络结构同样包含编码器与解码器模块. 其编码器部分参考了深度网络的架构设计,解码器部分则采用精简结构,仅包含 4 个卷积层,解码器网络将编码器中所提取的图像特征进行整合. 首先对图像特征实施降维处理,接着按行方向拼接特征矩阵,再执行多级卷积运算,将矩阵缩放 0.01,最终输出轴角矩阵和平移矩阵. 使用矩阵预测出相机位置变化的平移运动和旋转运动. 最终输出相邻单目图像帧之间的 6 自由度位姿变换参数.

2　双重注意力模块和损失函数

2.1　通道注意力模块

改进型通道注意力模块的设计源于 SENet[23] 的基础架构,其结构演进如图 4 所示. 通道注意力模块的输入为特征图 A,通过双路径特征聚合策略实现通道维度建模:首先对输入特征执行全局平均池化与全局最大池化操作,分别生成空间压缩特征向量 *AP* 和 *MP*,其数学表达如公式(4)、公式(5)所定义. 公式中 k 为特征通道索引,(i,j) 表示第 k 个通道特征图的空间位置坐标.

$$AP_k = \frac{1}{W \times H} \sum_{i,j}^{H,W} A_{ijk} \tag{4}$$

$$MP_k = max_{i \in H, j \in W} A_{ijk} \tag{5}$$

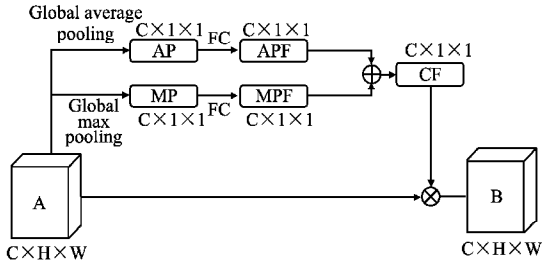

图 4　通道注意力模块

Fig. 4　Channel attention module

在保持 *AP* 和 *MP* 两个特征不变的前提下,通过全连接层对其实施非线性变换,随后采用 ReLU 激活函数生成激活后的特征向量 *APF* 与 *MPF*. 最终通过逐元素相加的方式将 *APF* 和 *MPF* 进行特征融合,形成综合表征向量,在综合表征向量上继续使用全连接层,通过 sigmoid 激活函数对输出值进行归一化处理,将其映射至[0,1]区间. 由此生成与特征图 A 通道维度一致的通道注意力权重向量,作为各通道的重要性评分. 将该分数乘到特征图 A 对应的通道上,从而获得通道加权的特征图 B. 该过程如公式(6)所示,其中运算符表示逐元素乘法(哈达玛积). 其中 * 为点乘.

$$B_{ij}^k = A_{ij}^k * CF^k \tag{6}$$

2.2　空间注意力模块

传统的卷积运算只具有局部感受野,提取的局部特征缺乏全局上下文信息,无法强调局部特征之间的联系. 这导致相同距离的像素在卷积后变得有些不同,这些差异引入了不连续性. 因此,在估计深度时,容易出现细长物体的中断等问题,并且物体边缘的模糊与实际情况不符. 为了使网络高效捕获场景全局上下文信息并充分建模特征间依赖关系,从而提高单目深度估计精度,本研究通过引入空间注意力机制优化网络架构. 空间注意力模块如图 5 所示,通过对每个位置的特征进行加权求和,选择性地聚合每个位置的特征以更新特征,使

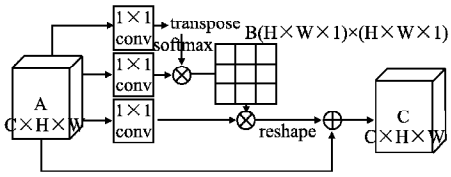

图 5　空间注意力模块

Fig. 5　Position attention module

得特征之间的全局相关性相互促进,从而获得需要关注的详细信息. 由于内窥镜场景都是人体组织,空间注意力模块可以将更广泛的背景信息编码到局部特征中,从而提高全局特征表示能力.

2.3　损失函数

2.3.1　光度重投影误差

本方法通过以下流程实现图像重建与误差评估:首先,基于源图像 D_s 的深度图数据,结合相机内参矩阵 K,依据透视投影逆变换原理将二维像素坐标反投影至三维空间坐标系;其次,通过相邻帧间位姿变换矩阵将三维点云映射至目标帧 $D_{s'}$ 的相机坐标系;随后,基于透视投影模型正向计算获得目标帧的二维投影坐标;最终,通过建立源-目标图像间的像素级对应关系,采用双线性采样机制提取颜色信息进行图像重建,并通过比对重建图像 $D_{s'\to s}$ 与原始图像的像素差异构建优化目标函数.

$$L_r = (1-\varepsilon) \parallel D_s - D_{s'\to s} \parallel \tag{7}$$

$$Z_s = cam(h_s, K, D_s) \tag{8}$$

$$D_{s'\to s} = D_{s'} < dep(Z_s, T_{s\to s'}, k) \tag{9}$$

Z_s 为被测图像 D_s 的空间三维坐标;h_s 为被测图像预测的深度;ε 为常数;$T_{s\to s'}$ 为被测图像与下一帧图像的相机位姿关系矩阵;$cam(\,)$ 为计算空间三维坐标的函数;< > 为采样运算符;$dep(\,)$ 为深度值计算函数.

2.3.2　多尺度结构相似性损失

SSIM 是一种经常用于图像处理任务的指标,用于比较重建图像和原始图像之间的结构信息,衡量两幅图像之间的相似程度. *SSIM* 考虑图像的 3 个关键特征:亮度(图像像素的平均值)、对比度(图像像素的方差)和结构(相关系数). *SSIM* 取值范围为[0,1],其值越大,两帧图像相似度越高. *MS_SSIM* 是 *SSIM* 的多尺度版本,它还考虑了图像的分辨率. *MS_SSIM* 比 *SSIM* 更好地保留了高频区域中的对比度(*MS_SSIM* 为模糊边界分配了更高的权重). 对应的损失函数计算公式为:

$$L_{ms} = \frac{\delta}{2}(1 - MS\_SSIM) \tag{10}$$

$$SSIM(p) = \frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1} \cdot \frac{2\sigma_{xy} + C_2}{\sigma_x^2 + \sigma_y^2 + C_2} = l(p) \cdot cs(p) \tag{11}$$

$$MS\_SSIM(p) = l_M^{\gamma}(p) \cdot \prod_{j=1}^{M} cs_j^{\beta_j}(p) \tag{12}$$

其中 p 为像素,μ_x、μ_y 为平均值,σ_x、σ_y 为方差,σ_{xy} 为两帧的

协方差，C_1、C_2 为稳定变量，$l(p)$为亮度相似度，$cs(p)$为简化后的对比度相似度和结构相似度，M 为尺度级. 为方便起见，本文设 $\gamma=\beta_j=1$，对于 $j=\{1,\cdots,M\}$.

2.3.3 边缘感知平滑性损失

由于光度损失对于弱纹理和均匀的内窥镜图像没有足够的信息，因此容易在存在边缘的地方引起不连续性. 本文引入了边缘感知平滑损失函数，边缘感知平滑损失函数通过自适应抑制不准确的预测信号，同时保持边缘区域的细节清晰度，从而生成更为平滑且结构保持完整的深度估计图. 其计算公式为：

$$L_s=|\alpha_s d_s^*|e^{-|\alpha_s D_x|}+|\alpha_y d_s^*|e^{-|\alpha_y D_x|} \tag{13}$$

式中 d_s^* 为视差图 d_s 的归一化计算，用于解决尺度不确定性问题. 为了防止训练目标陷入局部极小而导致空洞现象，减少伪影，本文引入了多尺度深度预测. 将较低分辨率的深度图上采样到输入图像的分辨率，然后重新投影、重新采样并计算该较高输入分辨率下的误差. 综上所述，本文的损失函数为：

$$L=\varphi L_P+\phi L_s=\varphi(L_r+L_{ms})+\phi L_s \tag{14}$$

其中 φ 和 ϕ 是相关损失函数的权重.

3 实验结果与分析

3.1 实验设置

本文基于 PyTorch 框架实现了内窥镜深度估计模型，并在一台配备了 NVIDIA GeForce RTX 3080 GPU 的机器上完成 30 轮迭代训练. 本文使用 Adam 优化器对深度网络与位姿网络进行联合训练，batch size 为 8，初始学习率为 0.0001，并每 5 个轮次减少 10%，数据集的输入与输出分辨率为 256×256dpi.

3.2 数据集

训练深度估计模型需要包含 RGB 图像及对应真实深度图的数据集，但因人体器官狭小和设备的限制，一般难以通过传感器获取人体胃肠道内窥镜的精准深度标注. 当前研究多采用合成数据集（含 3D 建模生成的 RGB 图像及对应深度图）进行评估，本文采用 Endoslam[21] 数据集进行实验评估. EndoSLAM 数据集专门为内窥镜的 SLAM 研究设计，它整合了 3D 点云、胶囊/传统内窥镜的视频以及合成数据，针对了 6 种猪器官的配准与检测任务. 这些数据通过机械臂、4 种内窥镜设备和 2 款高精度 3D 扫描仪，从 8 类猪胃肠道器官中采集获得. 该数据集包括 21887 张结肠图片，12558 张小肠图片和 1548 张胃图片，像素大小为 320×320dpi. 本文没有使用 Endoslam 数据集中的所有图像作为模型的数据集，而是从中挑选了 6238 幅内窥镜图像，其中 4502 幅用于训练集，1292 幅用于验证集，剩余 444 幅用于测试.

3.3 性能评价指标

实验的评估指标主要分为以下 7 个指标：*AvsRel*（绝对相对误差）、*SqRel*（平方相对误差）*RMSE*（均方根误差）、*RMSE log*（*log* 均方根误差）以及 $\delta<1.25$，$\delta<1.25^2$ 和 $\delta<1.25^3$ 这 3 类阈值. 其中 *Abs Rel*，*SqRel*，*RMSE* 和 *RMSE log* 指标表示误差，结果越小越好，$\delta<1.25$，$\delta<1.25^2$ 和 $\delta<1.25^3$ 这 3 项指标表示精度，结果越大越好，具体公式如下所示：

$$Abs\ Rel=\frac{1}{N}\sum\frac{|y_{pred}-y_{gt}|}{y_{gt}} \tag{15}$$

$$Sq\ Rel=\frac{1}{N}\sum\frac{\|y_{pred}-y_{gt}\|^2}{y_{gt}} \tag{16}$$

$$RMSE=\sqrt{\frac{1}{N}\sum\left\|\frac{1}{y_{pred}}-\frac{1}{y_{gt}}\right\|^2} \tag{17}$$

$$RMSE\ log=\sqrt{\frac{1}{N}\sum\|\log y_{pred}-\log y_{gt}\|^2} \tag{18}$$

$$\delta=MAX\left(\frac{y_{pred}}{y_{gt}},\frac{y_{gt}}{y_{ored}}\right) \tag{19}$$

其中 N 为样本数量，y_{gt} 以及 y_{pred} 分别为实际深度值和预估深度值，δ 表示实际深度与预估深度比例的最大值.

3.4 深度估计评估

为了定量评估本文所提出的内窥镜深度估计算法的性能，使用几种最先进的算法进行了比较分析. 在内窥镜场景下，有几篇文章评估了他们的模型[24,25]. 但是，它们都没有提供代码供参考. 在此基础上，本文选择了几种最先进的且适用于内窥镜下的深度估计方法 SfmLearner[26]，Monodepth2[20]，DepthAnything[27]，AF-SfM[28]，Packnet-Sfm[29] 和 EndoSfmLearner[19]. 基于 5 个评估指标的不同算法的定量性能评估结果如表 1 所示. 从表 1 中的值可以看出，本文算法的绝对相对误差（AbsRel）、平方相对误差（SqRel）均方根误差（RMSE）、均方根对数误差（RMSElog）、以及不同阈值下的准确率（δ）都取得了很不错的效果. 与 Monodepth2 相比，本文在绝对相对误差（AbsRel）、平方相对误差（SqRel）均方根误差（RMSE）、均方根对数误差（RMSElog）较于 Monodepth2 分别减少 9.8%、23.7%、20.5%、7.3%，在阈值为 $\delta<1.25$、$\delta<1.25^2$ 和 $\delta<1.25^3$ 的准确度分别提升 7.2%、5%、3.6%. 这一结果表明相较于其他网络模型，本模型在深度预测方面具有更高的清晰度和精确度.

表 1 不同深度估计算法的结果比较

Table 1 Comparison of results of different depth estimation algorithms

方法	AbsRel	SqRel	RMSE	RMSE log	$\delta<1.25$	$\delta<1.25^2$	$\delta<1.25^3$
Endo-Sfm	0.309	0.024	0.062	0.343	0.613	0.847	0.943
Monodepth2	0.324	0.026	0.068	0.352	0.601	0.810	0.919
DepthAnything	0.293	0.021	**0.051**	0.308	0.632	**0.856**	0.944
SfmLearner	0.315	0.028	0.065	0.352	0.607	0.832	0.935
AF-Sfm	0.308	0.023	0.060	0.337	0.619	0.849	0.937
Packnet-Sfm	0.302	0.024	0.065	0.341	0.622	0.833	0.945
Ours	**0.292**	**0.020**	0.054	**0.294**	**0.645**	0.852	**0.953**

在定量对比的基础上，为更直观地凸显本文方法的优势，图展示了本文方法与 4 种方法在深度预测上的对比结果，并进行了定性分析. 图 6 首列为原始 RGB 图像，随后 5 列依次为 Endo-Sfm、Monodepth2、Packnet-Sfm、SfmLearner 和本文算法对应的深度图. 通过观察图片，可以发现 Monodepth2 算法边缘处理能力较差，深度图模糊；Endo-Sfm 算法相比于 Sfmlearner 算法确实提高了性能，但是 Endo-Sfm 和 Sfmlearner 算法在处理数据集中具有深度估计算法难以处理的亮点时，效果较差. Packnet-Sfm 算法则在 Endoslam 数据集上存在明显的黑洞问题，难以判断深度信息；相比之下，本文的方法表现出更优越的整体性能，在物体的具体细节和整体完整度上取

得了更好的效果,在保留局部特征的同时,使物体的细节更加精确,最终生成的深度图像使物体轮廓更加清晰,边缘更加锐利. 因此,本文算法在深度图的主观视觉效果上优于其他自监督的算法.

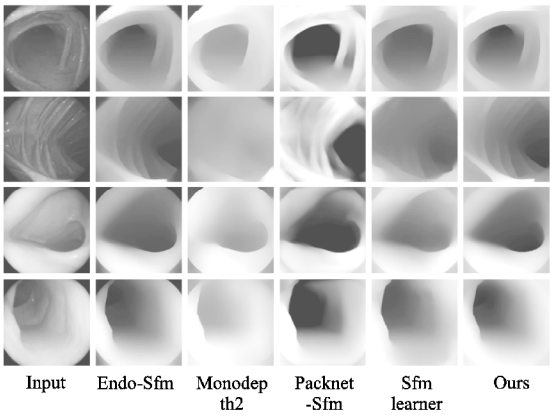

图 6　不同深度估计算法的结果比较

Fig. 6　Comparison of results from different depth estimation algorithms

3.5　消融实验

为了验证本文方法中引入通道注意力模块和空间注意力模块对模型性能的影响,本文进一步进行了消融实验. 其中 Baseline 表示 Monodepth 2 方法,Baseline + PA 表示增加空间注意力模块,Baseline + CA 表示增加通道注意力模块,而 Baseline + CA + PA 与本文的方法相比则是在串联处理次序上的不同. 从表 2 中的值可以看出,空间注意力模块的引入明显提高了模型的性能,但是通道注意力模块和空间注意力模块的单独引入时,模型的性能提升有限,而当引入双重注意力模块时在大多数指标上对模型性能的提高更大. 并且按照本文的串联次序效果略优于传统的方法,这是因为空间注意力模块和通道注意力模块联合使用时,通过跨层级特征融合机制,有效增强上下文感知能力,缓解单层特征表征能力受限的缺陷. 该设计提高了网络对上下文信息的利用,显著提升了网络的特征判别性. 从而通过各项指标凸显出嵌入双中注意力机制时深度估计的效果最佳. 图 7 为引入双重注意力机制前后的对比效果. 从图中可以清楚地看到,具有双重注意力机制的深度图具有更好的连续性和更清晰的边缘. 综上所述,引入双注意力机制确实为模型提供了显着的增益.

表 2　消融实验的结果

Table 2　Results of the ablation experiment

方　法	Backhone	AbsRel	SqRel	RMSE	RMSElog	$\delta<1.25$	$\delta<1.25^2$	$\delta<1.25^3$
Baseline	Resnet-18	0.324	0.026	0.068	0.352	0.601	0.810	0.919
Baseline	Resnet-50	0.320	0.024	0.061	0.345	0.614	0.823	0.927
Baseline + CA	Resnet-18	0.318	0.023	0.063	0.337	0.627	0.828	0.928
Baseline + PA	Resnet-18	0.302	0.021	0.061	0.324	0.634	0.835	0.936
Baseline + CA + PA	Resnet-18	0.294	0.021	0.058	0.303	0.641	0.842	0.947
Baseline + DA(ours)	Resnet-18	**0.292**	**0.020**	**0.054**	**0.294**	**0.645**	**0.852**	**0.953**

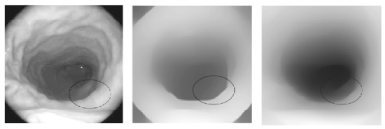

图 7　引入双重注意力机制后的前后比较

Fig. 7　Comparison before and after introducing dual attention mechanism

4　结　语

本文提出了一种“新的嵌入双重注意力机制的无监督单目内窥镜深度估计网络”模型,该模型主要包含两个核心部件,即通道注意力模块和空间注意力模块. 通过引入双重注意力机制来学习可区分和高纹理结构,从而获得场景结构的内涵和更为有效的特征表示,并有效地融合不同层次的特征. 本文还通过引入光度重投影误差和结构相似性和边缘感知平滑作为损失函数,来提高模型表示图像细节的能力,该损失函数可以更好地保留高频信息并更好地保持亮度和颜色不变性. 此外,通过一系列实验证明本文所提出的网络模型能够产生更清晰、更精确的深度预估,并在 Endoslam 数据集上取得了更为突出的表现.

References:

[1] Hsia C H, Chiang J S, Li H T, et al. A 3D endoscopic imaging system with content-adaptive filtering and hierarchical similarity analysis[J]. IEEE Sensors Journal, 2016, 16(11): 4521-4530.

[2] Mahmood F, Durrn J. Deep learning and conditional random fields-based depth estimation and topographical reconstruction from conventional endoscopy[J]. Medical Image Analysis, 2018, 48(13): 230-243.

[3] Pei L Y, Chun S H, Yu Q H, et al. Surgical navigation system based on the visual object tracking algorithm[C]//4th Annual International Conference on Network and Information Systems for Computers(ICNISC), 2018: 160-164.

[4] JIANG J J, LI Z Y, LIU X M. Deep learning based monocular depth estimation methods: a survey[J]. Chinese Journal of Computers, 2022, 45(6): 1276-1307.

[5] CIIEN Y F. Progress of visual depth estimation a-nd point cloud mapping[J]. Chinese Journal of Liquid Crystals and Displays, 2021, 36(6): 896-911.

[6] Nikolaus M, Eddy I, Philip H, et al. A large dataset-to train convolutional networks for disparity, optical flow, and scene flow estimation[C]//IEEE Conferenceon Computer Vision and Pattern Recognition(CVPR), 2016: 4040-4048.

[7] Pang J, Sun W, Ren J S, et al. Cascade residual learning: a two-stage convolutional neural network for stereo matching[C]//IEEE International Conference on Computer Vision Workshops(ICCVW), 2017: 878-886.

[8] Alex K, Martirosyan H, Dasgupta S, et al. End-to-end-learning of geometry and context for deep stereo regression[C]//IEEE International Conference on Computer Vision(ICCV),2017:66-75.

[9] Grasag O G, Bernal E, Casado S, et al. Visual SLAM for handheld monocular endoscope[J]. IEEE Transactions on Medical Imaging, 2013,33(1):135-146.

[10] Leonard S, Sinha A, Reite A, et al. Evaluation and stability analysis of video-based navigation system for functional endoscopic sinus surgery on in vivo clinical data[J]. IEEE Transactions on Medical Imaging,2018,37(10):2185-2195.

[11] WANG T M, ZHANG X H, ZHANG X B, et al. Review of research progress on laparoscopic augmented-reality navigation[J]. Robotics,2019,41(1):124-136.

[12] Qiu L, Ren H. Endoscope navigation and 3D reconstruction of oral cavity by visual SLAM with mitigated data scarcity[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops(CVPR),2018:2197-2204.

[13] Grigo R A, Jiang F, Rho S, et al. Depth estimation from single monocular images using deep hybrid network[J]. Multimedia Tools and Applications,2017,76(18):18585-18604.

[14] Chen S N, Tang M X, Kanjm, et al. Encoder decoderwith densely convolutional networks for monocular depth estimation[J]. Journal of the Optical Society of America A,2019,36(10):1709-1718.

[15] Liu X, Sinha A, Unberath M, et al. Self-supervised learning for dense depth estimation in monocular endoscopy[C]//OR 2.0 Context-Aware Operating Theaters, Computer-Assisted Robotic Endoscopy, Clinical Image-Based Procedures, and Skin Image Analysis,2018:128-138.

[16] Li Y. EndoDepthL: lightweight endoscopic monoculardepth estimation with CNN-transformer[C]//IEEE International Conference on Bioinformatics and Biomedicine(BIBM),2023:4344-4351.

[17] Liu S Y, Fan J F, Yang Y, et al. Monocular endoscopy images depth estimation with multi-scale residual fusion[J]. Computers in Biology and Medicine,2024,16(9):235-243.

[18] Yang Z Y, Pan J J, Dai J, et al. Self-supervised endoscopy depth estimation framework with CLIP-guidance segmentation[J]. Biomedical Signal Processing and Control,2024,9(5):132-140.

[19] Kustev B O, Guliz I G, Taylor L B, et al. EndoSLAM dataset and an unsupervised monocular visual odometryand depth estimation approach for endoscopic videos[J]. Med Image Anal, 2021, 7(13):1020-1028.

[20] Godard C, Mac Aodha O, Firman M, et al. Digging into self-supervised monocular depth estimation[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision,2019:3828-3838.

[21] 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.

[22] Godard C, Mac Aodha O, Brostow G J. Unsupervise-d monocular depth estimation with left-right consistency[C]//Proceedings of the IEEE Conference on Computer Visionand Pattern Recognition, 2017:270-279.

[23] Hu J, Shen L, Sun G. Squeeze-and-excitation networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,2018:7132-7141.

[24] Rau A, Edwards P, Ahmad O F, et al. Implicit domain adaptation with conditional generative adversarial networks for depth prediction in endoscopy[J]. International Journal of Computer Assisted Radiology and Surgery,2019,14(7):1167-1176.

[25] Hwang S J, Park S J, Kim G M, et al. Unsupervisedmonocular depth estimation for colonoscope system using feedback network [J]. Scnsors,2021,21(8):2691-2670.

[26] Borgli H, Thambawita V, Smedsrud P H, et al. HyperKvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy[J]. Scientific Data,2020,7(1):1-14.

[27] Yang L H, Kang B Y, Huang Z L, et al. Depth anything: unleashing the power of large-scale unlabeled data[C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR),2024: 10371-10381.

[28] Shao S W, Pei Z C, Chen W H, et al. Self-supervised monocular depth and ego-motion estimation in endoscopy: appearance flow to the rescue[J]. Med Image Anal,2022,7(8):102-112.

[29] Guizilini V, Ambrus R, Pillai S, et al. 3D packing forself-supervised monocular depth estimation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020: 2485-2494.

附中文参考文献:

[4] 江俊君,李震宇,刘贤明. 基于深度学习的单目深度估计方法综述[J]. 计算机学报,2022,45(6):1276-1307.

[5] 陈苑锋. 视觉深度估计与点云建图研究进展[J]. 液晶与显示, 2021,36(6):896-911.

[11] 王田苗,张晓会,张学斌,等. 腹腔镜增强现实导航的研究进展综述[J]. 机器人,2019,41(1):124-136.