SoftEnNet：面向结肠镜内机器人的协同单目深度估计与流明分割

Alwyn Mathew[1], Ludovic Magerand[2],Emanuele Trucco[2] 和 Luigi Manfredi[1*]

arXiv:2301.08157v1 [eess.IV] 19 Jan 2023

摘要—**结直肠癌是全球癌症死亡的第三大常见原因。光学结肠镜检查是检测结直肠癌的金标准；然而，检查过程中约有25% 的息肉会被漏诊。基于视觉的自主内机器人能够通过对结肠黏膜进行系统、全面的筛查，显著改善结肠镜检查流程。实现可靠机器人导航所需的三维环境理解和管腔跟踪，可为自主任务提供支持。我们提出了一种新颖的多任务模型，利用深度网络集成，同时预测稠密深度和管腔分割。深度估计子网络在视图合成的引导下以自监督方式进行训练；管腔分割子网络则采用监督方式训练。这两个子网络通过通路相互连接，从而实现信息交换并促进互学习。由于管腔位于图像视觉空间中最深处，管腔分割有助于估计最远位置的深度。反过来，估计深度也能指导管腔分割网络，因为管腔位置定义了场景中最远的位置。与其他环境不同，视图合成在结肠中常常会失效，原因包括可变形肠壁、无纹理表面、镜面反射以及宽视场图像畸变；这些挑战均由我们的处理流程加以应对。我们在合成数据集上进行了定性分析，并在结肠训练模型和真实结肠镜视频上进行了定量分析。实验表明，我们的模型能够从结肠镜图像中近乎实时地预测准确的尺度不变深度图和管腔分割。**

I. 引言

2020年，结直肠癌是全球报告病例数第三多的癌症，也是导致死亡人数第二多的癌症， [1] 在男性和女性中均占全球癌症发病数的10%，占所有癌症死亡人数的9.4%。根据对人口老龄化、人口增长和人类发展的预测，到 2040 [2]，全球结直肠癌新发病例数预计将达到320万。结直肠癌病例不断增加，且年轻一代的发病率持续上升，这对财政造成了沉重负担，也带来了公共卫生挑战。得益于更健康的生活方式，以及十年前实施的一项筛查计划，一些人类发展水平较高的国家结直肠癌发病率有所下降或趋于稳定。早期发现是预防转移、降低死亡率以及改善未来预后和生活质量的关键因素。这些统计分析凸显了健康生活方式对于降低风险因素的重要性、开展结直肠癌筛查项目以便早期发现，以及对结直肠癌预防和治疗进行新研究的重要性。

光学结肠镜检查是结直肠癌筛查的金标准，因为它兼具目视检查结肠黏膜和切除可能最终癌变的息肉这两项功能。这项检查过程复杂，会造成疼痛和不适。检查结果与操作者的技能有关 [3]。腺瘤检出率（ADR）是指内镜医师检查的所有患者中，至少检出一个结直肠腺瘤的患者所占比例。ADR 是评估操作技术技能相关检查质量的关键指标之一。自主内镜机器人能够提高检查质量，并降低检查结果对操作者技能的依赖。这类机器人是医疗器械 [4] ，能够在一定程度的自主性下对结肠黏膜进行系统筛查。机载视觉系统在导航中的关键任务是识别前进方向，并构建结肠壁的三维重建，以执行手术任务。准确性和快速处理能力是自主内镜机器人 [5]实现实时控制的关键。

本研究提出了一种用于结肠镜检查的实时单目深度估计与管腔分割系统，可为实现自主运动和手术任务铺平道路。当结肠内部空间有限、难以使用三维立体视觉时，单目相机能够满足这一限制。与此前报道的工作相比，本研究的优势包括：

1）一种新颖的多任务网络 SoftEnNet，可同时预测用于内窥镜机器人导航的稠密深度估计结果并进行管腔分割。2）多任务网络 SoftEnNet 中深度与管腔子网络之间的新颖协同架构，可使两个任务均受益。3）一种新颖的有效性掩码，可在损失计算中排除不参与视图合成的像素，这些像素受到镜面光照或低纹理表面的影响，或不满足朗伯假设。

II. 相关工作

A. 深度估计

过去几年中，基于视频的单目稠密深度估计取得了显著进展，但结肠镜检查领域却并非如此。由于结肠表面纹理稀疏、易变形且具有反光性，逐像素深度估计在结肠镜检查中仍是一个尚未解决的问题。由于结肠尺寸的限制，结肠镜无法

本研究得到英国工程与物理科学研究理事会（EPSRC）资助，项目编号为 EP/W00433X/1[1]Alwyn Mathew 和 Luigi Manfredi 任职于英国邓迪大学医学院影像科学与技术部。

*l.manfredi@dundee.ac.uk[2]Ludovic Magerand 和 Emanuele Trucco 任职于英国邓迪大学科学与工程学院计算学科。

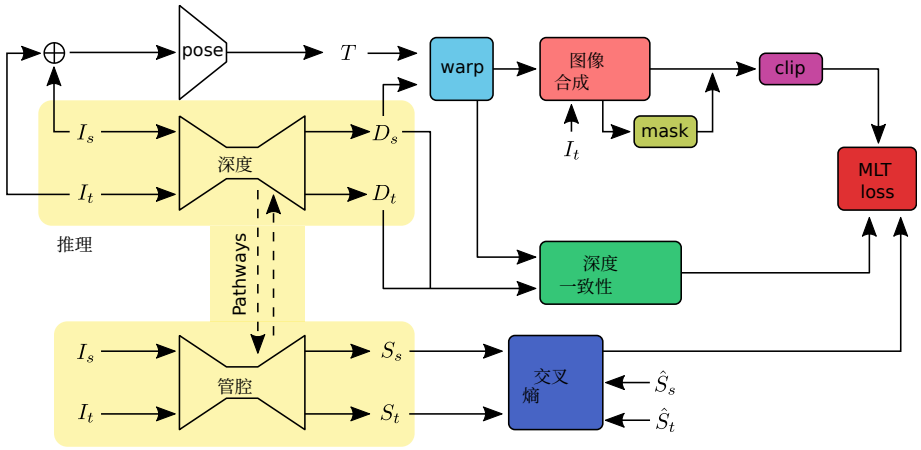

图 1. **多任务训练流程**：输入图像（I_s, I_t）是来自同一视频序列的连续帧。深度 子网络预测独立的逐像素稠密深度图（D_s, D_t），分别对应源 I_s 图像和目标 I_t 图像。姿态 子网络以拼接后的图像序列（I_s, I_t）为输入，并预测旋转 R 和平移 t，以定义源图像与目标图像之间的变换 T。变形 函数利用预测的深度和姿态，将源图像坐标变形至目标图像坐标，反之亦然。图像合成 通过变形后的图像坐标，将源图像 I_s 投影至目标图像 I_t。将合成图像与真实图像进行比较，以构造损失。使用裁剪 函数去除图像合成损失中的异常值。管腔子网络在源 I_s 图像和目标 I_t 图像中定位管腔（S_s, S_t）。通路在任务网络之间共享相关特征。使用交叉熵损失，以管腔分割真值（$\hat{}S_s$, $\hat{}S_t$）指导管腔解码器头。*MLT* 对任务损失进行加权，以强制实现同方差任务不确定性。 $\oplus$ 表示图像拼接。在推理阶段（黄色），仅使用深度与管腔子网络。

携带额外的深度传感器来采集深度图真值。结肠易变形的特性也使得通过计算机断层扫描（CT）获取的数据并不可靠。此前已有研究探索利用结肠建模线索和特征几何来实现结肠三维重建。Hong 等人 [6] 利用结肠皱襞轮廓估计深度。Zhao 等人 [7] 结合运动恢复结构（SfM）和明暗恢复形状（SfS）技术重建结肠表面，但深度估计结果不一致。研究人员利用色素内镜检查获得的图像增强结肠表面纹理，以改进SfM [8], [9], [10]。由于色素内镜检查并不常见且计算成本高，这些技术的应用有限。Ma 等人 [11] 采用SfM方法生成伪深度真值数据，并用这些数据训练深度学习网络来预测深度。Ma等人 [12] 提出了一种基于视频的自监督深度估计方法，并约束帧间的时间一致性。由照片级真实感仿体和CT扫描生成的合成数据已得到广泛研究，以助力数据驱动方法 [13],[14]。这些方法使用渲染图像进行训练，而非真实的结肠镜图像，因此在真实数据上的表现较差。

B. 语义分割

大多数结肠语义分割方法以息肉为目标，而我们的分割子网络专注于管腔分割，这对内镜机器人导航至关重要。早期的尝试大多采用传统图像处理技术，如分水岭算法 [15]，霍夫变换 [16] 和 Canny 边缘检测 [17]。Wang 等人 [18] 提出了基于 SegNet [19] 的实时息肉分割方法，以及全卷积膨胀网络。Jha 等人 [20] 提出了 ResUNet++，其分割效果优于 UNet 架构。Wang等人 [21] 提出了一种边界感知神经网络，Jha 等人 [22] 则提出了用于分割的 DoubleUNet 架构。Huang等人 [23] 采用低内存延迟的 HarDNet 骨干网络 [24]，结合级联部分解码器，实现快速且准确的息肉分割。Tomar 等人 [25]采用基于迭代细化方法的硬注意力，并使用大津阈值法。

III. 方法

本研究通过一种新颖的多任务网络（图 1），同时估计结肠镜检查 RGB图像流中的深度并分割流明。该网络由三个子网络组成：深度、姿态和管腔。自监督深度子网络由图像帧之间的视图合成损失进行引导。它以单目图像为输入，并输出稠密深度图。姿态子网络预测两个输入图像帧之间的刚性变换 $T_{t\to s}$。预测的深度 D_t 以及变换 $T_{t\to s}$ 可通过几何投影从源图像 I_s 合成目标图像 I_t。管腔子网络以监督方式预测管腔分割。该子网络以单幅图像为输入，并对图像中的管腔进行分割。连接深度子网络和管腔子网络的通路实现了互学习，并将已有技能结合起来，从而更快、更有效地学习新任务。推理时仅使用深度子网络和管腔子网络。

A. 自监督深度估计

我们的自监督单目深度估计框架用于所提出的深度估计框架中

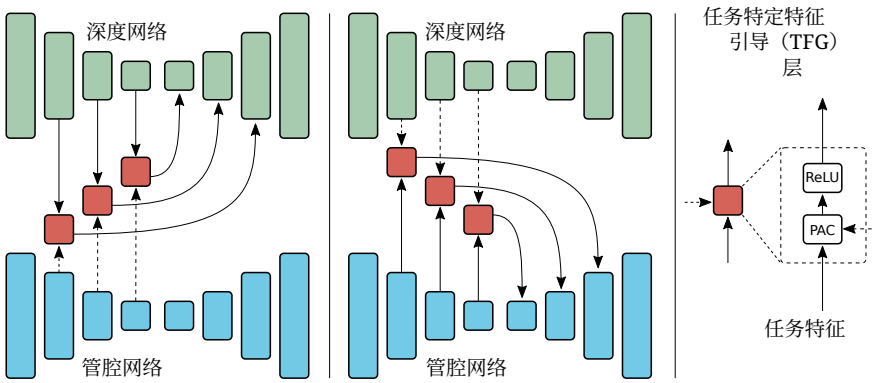

图 2. **SoftEnNet**：该多任务网络可同时预测图像中的深度和管腔位置。SoftEnNet 由针对每项任务的两个子网络组成。这些子网络通过任务特定特征引导层（*TFG*）连接，从而在共享功能特征空间的网络之间建立多条通路。引导层由像素自适应卷积层（*PAC*）和整流线性单元（*ReLU*）组成。引导特征用于调整作用于任务特征的*PAC* 核。左：从管腔网络到深度网络的通路。中：从深度网络到管腔网络的通路。右：任务特定特征引导（TFG）层。

该框架包含两个网络：

- 深度网络：N_D ： $I_t \rightarrow D_t$ 接收一幅单目图像 I_t 并预测相应的稠密深度图D_t。
- 姿态网络：N_P ： $(I_t, I_s) \rightarrow T_{t\rightarrow s}$ 接收连续的两帧图像（I_t, I_s），并预测源 I_s 图像与目标 I_t 图像之间六个自由度（DOF）的刚性变换 $T_{t\rightarrow s}$ 。

自监督深度通过利用视频帧之间的视图合成[26] 作为监督来实现。下面我们将介绍图像合成函数及其逆函数（π， π^{-1}）、图像重建损失 L_v、深度一致性损失 L_c、边缘引导的平滑损失 L_e和有效掩码 V，这些用于自监督深度和姿态损失。

1）图像合成：投影函数 π 将三维点云映射到二维图像坐标，反投影函数 π^{-1} 则将二维图像映射到三维世界坐标。通过刚性变换 $T_{t\rightarrow s}$、投影 π 和反投影π^{-1} 函数，将目标图像 I_t 的图像坐标投影到源图像 I_s，然后使用重映射函数 ζ对像素值进行逆向采样，即可合成目标图像。可根据相应输入图像 I_t 的估计深度 D_t 生成点云 P_t，如下所示：

$$p_s = \pi(T_{t\rightarrow s}P_t), \quad P_t = \pi^{-1}(p_t, D_t), \quad \hat{I}_t^{ij} = \zeta(I_s^{i'j'}) \tag{1}$$

其中， p_t 是图像坐标集合， P_t 是与目标图像 I_t对应的三维点云。姿态网络估计从目标图像到源图像的相对姿态 $T_{t\rightarrow s}$ 。该相对姿态可用于将目标点云转换为源点云 $P_s = T_{t\rightarrow s}P_t$。利用投影模型 π，三维点云 P_s 可映射到源图像坐标 p_s ，如下所示。通过对源图像坐标进行逆向变形，并使用 ζ对 [27] 进行采样，可合成目标图像。投影后的源图像坐标 $p_s(i', j')$将用于采样像素 $\hat{}$ 值，并通过采样函数 ζ 生成目标图像 I_t。可微空间变换网络 [28] 采用双线性插值来合成图像。由于逆变换后的网格可能不会落在图像网格点上，因此使用插值来查找源图像中的近似像素值。

2）图像重建损失：参考帧与重建帧之间的光度损失 L_v作为深度网络的自监督损失。遵循以往研究 [29], [30], [31], ，我们将结构相似性（SSIM） [32]与 L1 结合用于光度损失，权重为 η。这样既利用了 SSIM（记作 S。）处理复杂光照变化的能力，也利用了 L1 损失对异常值的鲁棒性，具体如下：

$$L_v^s(I_t, \hat{I}_t) = \frac{1}{|V|}\sum_{p\in V}\eta\frac{1 - S(I_t(p), \hat{I}_t(p))}{2} + (1-\eta)|I_t(p) - \hat{I}_t(p)| \tag{2}$$

这里， $|V|$ 是有效性掩码V中的有效像素数量。屏蔽违反非零光流和朗伯表面假设的 V 像素，可以避免受损梯度传播到网络中。有效性掩码的计算方法如下：

$$V = M_{ego} \cdot (|I_t(p) - \hat{I}_t(p)| < |I_t(p) - I_s(p)|) \tag{3}$$

M_{ego} 是有效像素的二值掩码，与 [30] 类似。对于源图像和目标图像之间因缺乏自运动、特殊照明、镜头伪影和低纹理而看起来静止的像素，在 V 中将其标记为无效。违反假设的光度损失部分会产生较大的损失，从而导致较大的梯度，可能使性能变差。按照 [33]，我们使用 θ 对损失进行裁剪，其中 θ 是 L_v 的 p^{th} 百分位数，以移除异常值，如 $L_v(s_i) = min(s_i, \theta)$，其中 s_i 是 L_v 中的 i^{th} 代价。

表 I SoftEnNet 在 UCL 2019 合成数据上的定量研究。在简单路径模型中，TFG 层中的 PAC 被卷积替换。最佳值以粗体标出，次佳值以下划线标出。

模型	深度							语义	
	绝对相对误差 ↓	平方相对误差 ↓	均方根误差 ↓	对数均方根误差 ↓	$\delta < 1.25 \uparrow$	$\delta < 1.25^2 \uparrow$	$\delta < 1.25^3 \uparrow$	流明 IoU ↑	Mean 交并比 ↑
SfMLearner [29]	0.451	0.711	1.768	0.537	0.366	0.618	0.785	-	-
单目深度1 [27]	0.444	0.752	1.755	0.531	0.371	0.625	0.790	-	-
DDVO [34]	0.452	0.772	1.768	0.538	0.366	0.618	0.785	-	-
Monodepth2 [35]	0.446	0.756	1.757	0.532	0.360	0.624	0.789	-	-
高分辨率深度 [36]	0.448	0.762	1.762	0.534	0.369	0.621	0.788	-	-
AF-SfMLearner [37]	0.387	0.618	1.648	0.480	0.420	0.686	0.831	-	-
AF-SfMLearner2 [38]	0.352	0.545	1.581	0.448	0.445	0.712	0.852	-	-
基线深度	0.332	0.565	1.364	0.376	0.497	0.789	0.913	-	-
基线流明	-	-	-	-	-	-	-	0.912	0.954
简单路径	0.150	0.144	0.720	0.187	0.814	0.947	0.981	**0.931**	**0.964**
SoftEnNet	**0.145**	**0.112**	**0.650**	**0.184**	**0.823**	**0.954**	**0.983**	**0.931**	**0.964**

3）深度一致性损失：深度一致性 [39] 损失 L_c确保深度网络利用源图像和目标图像 I_s，I_t 估计得到的深度图 D_s，D_t 表示同一个三维场景。最小化该损失不仅能促进序列中的深度一致性，也能促进由重叠图像批次构成的整个序列中的深度一致性。深度一致性损失定义如下：

$$L_c^s(D_s, D_t) = \frac{1}{|V|}\sum_{p\in V} DC^s$$
$$DC^s = \frac{|\hat{D}_s^t(p) - \hat{D}_s(p)|}{\hat{D}_s^t(p) + \hat{D}_s(p)} \tag{4}$$

其中，$\hat{}\ D_s^t$ 是具有姿态 $T_{t\to s}$ 的 D_t 的投影深度，对应于图像 I_s 中的深度；$\hat{}\ D_s^t$ 需要与 D_s 进行比较，但由于投影$\hat{}$不在 D_s的网格内，因此无法直接比较。因此，使用 D_s 上的可微双线性插值生成 D_s，并将其与$\hat{}\ D_s^t$进行比较。 DC 的倒数将作为遮挡掩码M_o，用于屏蔽从源视图到目标视图的遮挡像素，这些像素主要位于结肠结肠袋皱襞周围。图像重建损失 L_s 使用 M_o 掩码重新加权，如下所示：

$$L_v^{s\prime} = \frac{1}{|V|}\sum_{p\in V} M_o^s \cdot L_v^s$$
$$M_o^s = 1 - DC^s \tag{5}$$

4) 边缘引导平滑损失：由于均匀区域和低纹理区域无法为图像重建损失提供信息，因此使用边缘引导平滑损失对深度图进行正则化。该平滑度损失改编自 [34]，如下所示：

$$L_e^s(I_t, D_t) = \sum_{p\in I} e^{-\nabla I_t(p)} \nabla D_t'(p) \tag{6}$$

其中，∇ 表示一阶梯度，D_t' =(1/D_t)/$mean$(1/D_t)表示归一化逆深度。为避免优化器陷入局部极小值，深度估计和损失在多个尺度[29], [35]上计算。总损失在各尺度（S=4）和图像批次上取平均：

$$L_d = \frac{1}{S}\sum_{s=1}^{S} \sigma_1 L_v^{s\prime} + \sigma_2 L_c^s + \sigma_3 L_e^s \tag{7}$$

B. 管腔分割

管腔分割被定义为逐像素的类别分配。每个像素都会被分配一个类别标签 $c \in C = \{0, 1, 2\}$，其中 0 表示结肠壁，1 表示管腔。管腔分割网络 $N_L: I_t \to S_t$ 采用真实管腔图 S_t，S_s进行监督训练。该网络预测像素 S 属于类别 $c \in C$的后验概率，并在交叉熵损失中将其与独热编码的真实标注 S 进行比较，如下所示：

$$L_l = -\sum_{c\in C} \hat{S}_{t,c} \cdot log(S_{t,c}) + \hat{S}_{s,c} \cdot log(S_{s,c}) \tag{8}$$

C. 多任务联合优化

多任务学习涉及针对多个目标优化模型的问题。结合多目标损失的一种简单方法，是对每个任务的损失进行加权线性求和：

$$L = w_d L_d + w_l L_l \tag{9}$$

其中，w_d 和 w_l 分别是深度估计和管腔分割任务对应的权重。然而，性能容易受到权重值 w 的影响，而且这些超参数的调优成本很高。我们采用一种更便捷的方法来学习最优权重：遵循 [40]，通过考虑每个任务的同方差不确定性来加权多任务目标：

$$L = \frac{1}{2\gamma_1^2} L_d + \frac{1}{2\gamma_2^2} L_l + log(1+\gamma_1) + log(1+\gamma_2) \tag{10}$$

同方差不确定性 [40] 是特定于任务的，不会随输入变化而改变。增大噪声参数会降低相应任务的权重，该权重由可学习参数 γ决定。该目标会优化较大的不确定性，这应会使该任务的损失对总损失的贡献较小。这样，我们便能同时学习深度估计和管腔估计中具有不同单位或尺度的量。噪声参数 γ_1 和 γ_2 分别对深度和管腔进行加权。在实验中我们发现， γ_1 明显小于 γ_2。

IV. 网络架构

我们提出了一种新颖的协同多任务 SoftEnNet，可同时预测深度、姿态和管腔分割。SoftEnNet 由三个子网络组成：深度、姿态和流明。由于深度估计和管腔分割等相关任务能够相互受益，深度网络和流明网络通过通路连接，以促进信息交换，如图 2 所示。图像中视觉空间最远处的深度被定义为流明。稠密深度估计将提升流明预测效果，而管腔分割则有助于改善最远点处的深度预测。子网络之间建立了多尺度通路连接，使任务特定特征能够在由粗到细的尺度上进行交换。

我们在每个通路连接中定义了任务特定特征引导（TFG）层，将深度信息注入流明网络，并将管腔分割信息注入深度网络，如图 2 所示。受 [41],启发，TFG 层利用像素自适应卷积[42] （PAC），将引导任务的知识传递至目标任务网络。PAC 层解决了标准卷积操作固有的一些局限性，例如平移不变性使其与内容无关。深度和流明通路中的 TFG 层分别生成深度感知语义和语义感知深度特征表示。这些引导特征图消除了语义特征与深度特征之间的歧义，从而为预测提供了更好的特征空间。TFG 层由一个 PAC 层和后接的 ReLU [43]组成。为确保后续训练有更好的起始条件，我们使用常数 PAC-α对引导特征进行缩放。在初始训练轮次中，引导特征会乘以一个较小的常数，使其行为与传统卷积层非常相似。随后在训练过程中，PAC-α 逐渐增大至 1。

该深度子网络 受 [35] 启发，以 ResNet18[44] 作为骨干网络，并采用多尺度解码器，在四种不同尺度上预测深度。管腔子网络 采用与深度网络类似的 U-Net 架构，以 ResNet18 作为骨干网络，并通过解码器在最高尺度上预测管腔分割。遵循 [35], , 位姿子网络 与深度编码器共享权重，其解码器由一系列四个卷积层组成。

表 II UCL 2019 合成数据集上任务权重的消融研究。深度 和流明 分别表示基线深度网络和管腔分割网络。L 表示学习得到的任务权重。

模型	w_d, w_l	深度		语义	
		均方根误差 ↓	$\delta < 1.25$ ↑	流明 交并比 ↑	Mean 交并比 ↑
深度	-	0.681	0.811	-	-
流明	-	-	-	0.912	0.954
SoftEnNet	0.4, 0.6	0.679	0.815	0.901	0.948
SoftEnNet	0.6, 0.4	0.661	0.814	0.919	0.958
SoftEnNet	0.7, 0.3	0.669	0.819	0.920	0.958
SoftEnNet	1.0, 1.0	0.681	0.813	0.917	0.956
SoftEnNet	L	**0.650**	**0.823**	**0.931**	**0.964**

表 III

UCL 2019 合成数据上的 TFG PAC-α 消融研究。$0.0001 \rightarrow 1$ 表示 PAC-α 在训练期间从 0.0001 增加到 1。

像素自适应卷积-α	深度		流明	
	均方根误差 ↓	$\delta < 1.25$ ↑	流明 交并比 ↑	Mean 交并比 ↑
0.0001	0.685	0.816	0.935	0.966
0.001	0.658	0.821	0.936	0.966
0.01	0.646	0.820	0.936	0.966
0.1	0.888	0.811	0.928	0.962
1.0	0.761	0.788	0.928	0.963
$0.0001 \rightarrow 1$	0.650	0.823	0.931	0.964

V. 实验

A. 实现细节

模型使用 Pytorch [45] 和 Adam [46]优化器构建，以最小化训练误差。模型在 Nvidia A6000 上训练 40 个训练轮次，批量大小为 20。模型的初始学习率为 0.0001，并在 30 个训练轮次后减半。深度网络的 Sigmoid 输出 d 按 $D = 1/(d \cdot 1/(d_{max} - d_{min}) + 1/d_{min})$ 缩放，其中 $d_{min} = 0$ 和 $d_{max} = 20.0$ 单位对应于 $0 - 20cm$。UCL 2019 合成数据集 [47]图像网络的输入图像尺寸为 256×256 （高度 × 宽度），我们的 CTM 数据集则为 512×768。这些超参数通过经验确定： $\eta = 0.85$、 $\sigma_1 = 1.0$、 $\sigma_2 = 0.5$和$\sigma_3 = 0.001$。我们对 UCL 2019 数据集中的 50% 图像进行水平翻转。由于广角相机的主点偏离中心，因此不对 CTM 数据集进行图像翻转。截断损失 $p = 95$遵循 [33]。PAC-α 的初始值设为 0.0001，并每 5 个训练轮次翻倍，直至达到 1。

1) UCL 2019 合成数据集：UCL 合成数据[47] 使用人体 CT 结肠成像、人工分割和网格化生成。使用 Unity 引擎生成尺寸为 256×256 的 RGB 图像，采用虚拟针孔相机（90° 视场，配有两个光源），并生成最大范围为 20 cm 的深度图。该数据集包含 15,776 张 RGB 图像及深度图真值，按 8:1:1 的比例划分为训练集、验证集和测试集。数据集分为九个子集，每个子集包含三种

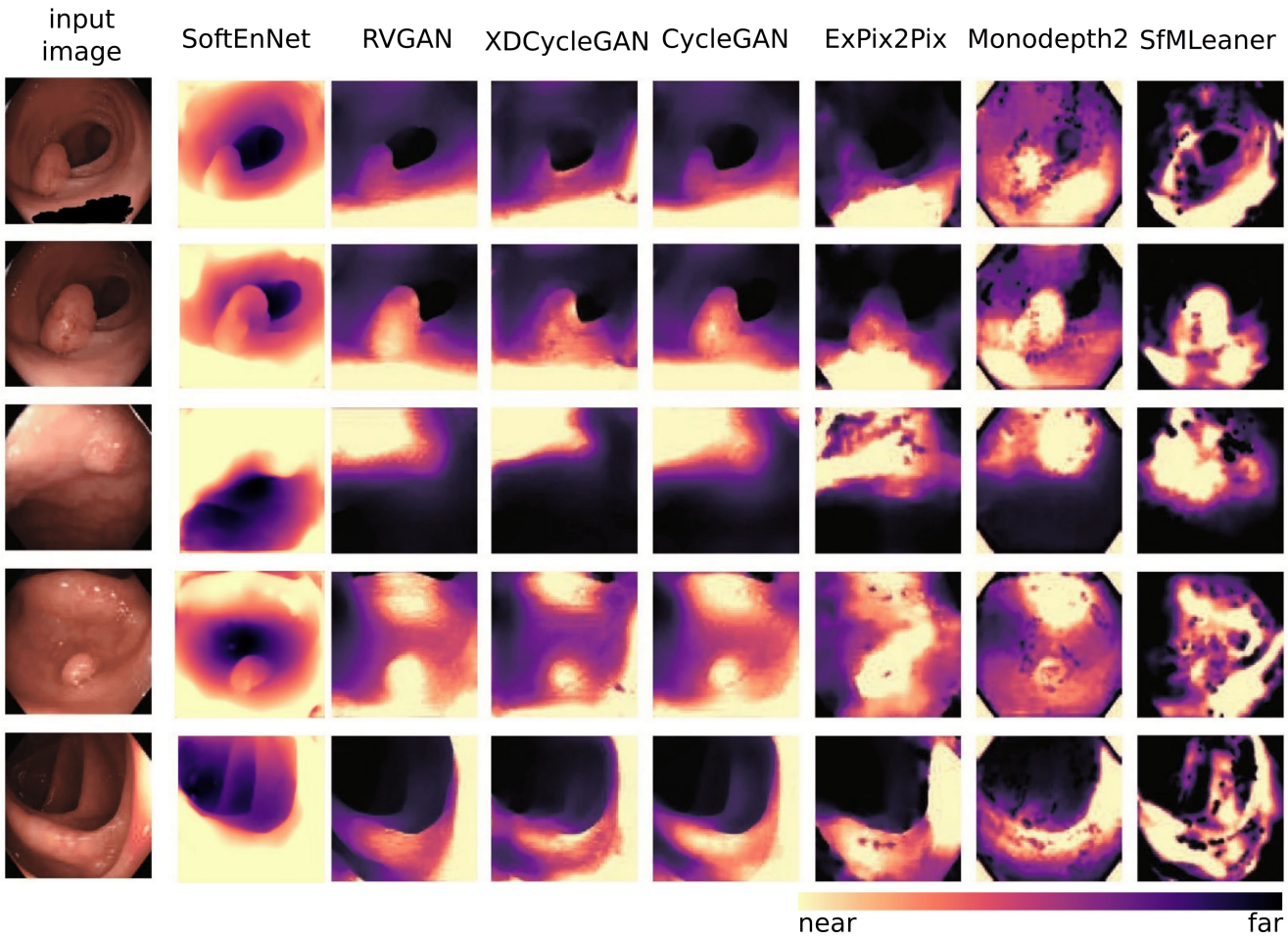

图 3. 在 RVGAN 真实结肠镜图像上对**SoftEnNet** 与 RVGAN [48], XDCycleGAN [49],CycleGAN [50], ExPix2Pix [47],Monodepth2 [35] 和 SfMLearner [29] 进行定性比较。

种光照条件和三种材料。光源在光束角、范围、颜色和强度方面各不相同，而材料则在颜色、反射率和平滑度方面各有差异。通过将深度图裁剪至第95百分位深度，从深度图真值中提取管腔分割真实标注。

2）结肠镜训练模型数据集：结肠训练模型（CTM）数据集是在用于医学专业人员培训的塑料模型内部，使用一台现成的全高清摄像机（1920×1080，30Hz，$132°$ 水平视场角，$65°$垂直视场角，$158°$ 对角视场角，焦距 45-20-70，摄像机周围设有白色环形灯）采集的。该模型模拟了人体结肠的 1:1解剖结构，包括内径、总长度和结肠袋皱襞，呈现出光学结肠镜检查中所见的情况。录制视频时，摄像机安装在线性执行器上，以约 2 cm/s 的恒定速度前后移动。该数据集包含约 10,000 张图像，按 8:2 的比例划分为训练集和验证集，涵盖不同的内外部光照设置，但未提供深度图真值。管腔分割的伪真实标注取自基线深度，并通过第95百分位深度进行裁剪。

B. 定量研究

真实结肠独特的结构限制，使得无法使用真正的深度传感器采集深度真实标注。因此，定量深度评估只能在 UCL 2019 等合成数据上进行，尽管当前合成数据的生成方式还不够逼真，无法模拟真实结肠表面的特性。SoftEnNet 的深度估计在 UCL 2019 测试数据集上使用 [53] 中的评估指标进行评估，并在表 I 中与其他最先进的方法进行比较。UCL 合成数据包含一些纹理，例如血管和结肠袋皱襞。深度评估表明，有效性掩码能够忽略不影响光度损失的像素，从而带来益处。研究还发现，裁剪掉高光度损失可提升模型性能。

C. 定性分析

难以对 Endomapper [51],LDPolypVideo[52],或 RVGAN [48] 数据集等真实结肠镜数据的深度进行定量评估。结肠的许多真实世界特征，例如不同组织类型的表面反射和光扩散，都难以在

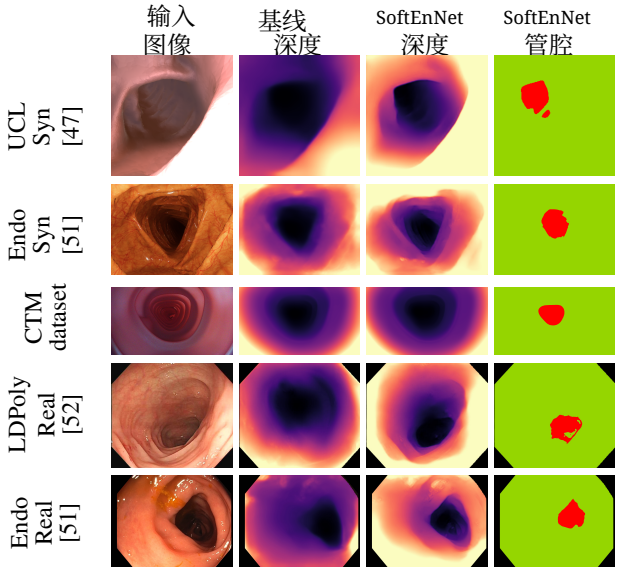

图 4. 对合成数据集（如 UCL 2019 合成数据集和 Endomapper 合成数据集）、结肠镜训练模型（CTM）数据集，以及真实结肠镜图像（如 LDPolyVideo 和 Endomapper 数据集）进行的定性分析（从上到下）。基线深度表示仅由深度网络估计的深度，SoftEnNet 深度表示由 SoftEnNet 估计的深度。SoftEnNet 管腔显示管腔分割结果（红色为管腔，绿色为结肠壁）。

模拟器中进行仿真。此外，真实结肠并非刚性结构，这使视图合成变得更加复杂。研究发现，仅使用图像重建损失在低纹理 CTM 上训练时，深度网络生成的深度图中会出现空洞。低纹理区域在真实结肠壁中十分常见；如何处理这些区域，对于依赖视图合成的数据驱动方法至关重要。第 III-A.2 节所述的有效性掩码能够有效去除镜面光照、镜头伪影和光流为零的区域。第 III-A.3 节所述的深度一致性损失有助于网络学习帧间几何一致性，并防止深度图出现闪烁。在 CTM 上训练的深度网络通过第 III-A.1 节的图像合成器，使用 Double Sphere [54] 等具有大视场的视场投影模型，学习直接从高度畸变的图像中预测深度。图 4 中对所提出的 SoftEnNet 进行的定性分析清楚地表明，这种协同关系能够呈现更多深度估计细节，尤其是结肠袋皱襞处的细节。图 3 展示了所提出模型与最先进的方法在真实结肠镜数据集 [48]上的深度预测定性比较。面对复杂的真实结肠镜图像时，所提出模型的鲁棒性优于 RVGAN [48],XDCycleGAN[49], CycleGAN [50], ExPix2Pix [47],Monodepth2 [35] 和 SfMLearner [29] 。

D. 消融研究

表 II 展示了考虑同方差不确定性时学习得到的任务权重与手工设计的任务权重之间的影响。表 III 展示了在训练过程中将 PAC-α 从较小的初始值逐渐增大，而非将 alpha 设为固定值的优势。较低的 PAC-α 表示子网络之间传递的信息较少，但在训练初期使用 PAC-$\alpha = 1$ 阶段会降低深度和管腔分割的准确率。

VI. 结论

理解三维结构并进行管腔跟踪，对于自主结肠镜内机器人的运动和手术任务至关重要。在本研究中，我们开发了一种新颖的协同网络 SoftEnNet，可同时预测深度并进行管腔分割。SoftEnNet 中的深度与管腔子网络利用另一项任务提供的引导来完成目标任务。我们还提出了一种新颖的有效性掩码，用于去除受镜面光照、低纹理表面以及不符合朗伯表面的影响的像素。本研究可用于实现自主任务，这些任务要求视觉系统与设备之间配备实时闭环控制器。本研究提出的视觉系统能够为自主运动算法提供运动方向，即管腔中心。未来的工作计划将当前视觉系统与运动 [5] 和低级控制器[55]集成。这包括测试输出在真实相机运动下的稳定性，因为相机运动会影响照明（由机器人携带），进而影响图像外观，并可能降低视频质量以及估计深度图和管腔位置的质量。

参考文献

[1] Y. Xi 和 P. Xu，“2020年全球结直肠癌负担及至2040年的预测”，转化肿瘤学，第14卷，第10期，第101174页，2021年。[2] H. Sung、J. Ferlay、R. L. Siegel、M. Laversanne、I. Soerjomataram、A. Jemal 和 F. Bray，“2020年全球癌症统计：Globocan对185个国家36种癌症发病率和死亡率的估算”，《*CA*：临床医师癌症杂志》，第71卷，第3期，第209–249页，2021年。[3] N. N. Baxter、M. A. Goldwasser、L. F. Paszat、R. Saskin、D. R. Urbach 和 L. Rabeneck，“结肠镜检查与结直肠癌死亡之间的关联”，《内科学年鉴》，第150卷，第1期，第1–8页，2009年。[4] L. Manfredi，“结肠镜检查内机器人：设计挑战与现有技术”，《机器人与人工智能前沿》，第209页，2021年。[5] L. Manfredi、E. Capoccia、G. Ciuti 和 A. Cuschieri，“用于结肠镜检查的软体气动双气囊尺蠖机器人（SPID）”，《科学报告》，第9卷，第1期，第1–9页，2019年。[6] D. Hong、W. Tavanapong、J. Wong、J. Oh 和 P. C. De Groen，“基于结肠镜图像的虚拟结肠结构三维重建”，《计算机化医学成像与图形学》，第38卷，第1期，第22–33页，2014年。[7] Q. Zhao、T. Price、S. Pizer、M. Niethammer、R. Alterovitz 和 J. Rosenman，“内窥镜图谱：由内窥镜视频帧重建的三维模型”，载于医学图像计算与计算机辅助干预国际会议。施普林格，2016年，第439–447页。[8] A. R. Widya、Y. Monno、M. Okutomi、S. Suzuki、T. Gotoda 和 K. Miki，“基于虚拟色素内镜图像生成的胃部三维重建”，载于*2020*年*IEEE*医学与生物学工程学会第*42*届年度国际会议（*EMBC*）。IEEE，2020年，第1848–1852页。[9] ——，“基于单目内窥镜视频的全胃三维重建与帧定位”，*IEEE*健康与医学领域转化工程期刊，第7卷，第1–10页，2019年。[10] ——，“使用GAN增强图像进行胃肠内窥镜检查中的自监督单目深度估计”，载于医学成像*2021*：图像处理，第11596卷。SPIE，2021年，第319–328页。

[11] R. Ma、R. Wang、S. Pizer、J. Rosenman、S. K. McGill 和 J.-M. Frahm，“用于确定缺失区域的结肠镜表面实时三维重建”，载于医学图像计算与计算机辅助干预国际会议。施普林格，2019年，第573–582页。[12] D. Freedman、Y. Blau、L. Katzir、A. Aides、I. Shimshoni、D. Veikherman、T. Golany、A. Gordon、G. Corrado、Y. Matias 等，“检测结肠镜检查中的覆盖不足”，IEEE医学影像汇刊，第39卷，第11期，第3451–3462页，2020年。[13] F. Mahmood 和 N. J. Durr，“基于深度学习和条件随机场的常规内窥镜检查深度估计与地形重建”，医学图像分析，第48卷，第230–243页，2018年。[14] M. Visentini-Scarzanella、T. Sugiura、T. Kaneko 和 S. Koto，“用于支气管镜检查辅助导航的深度单目三维重建”，国际计算机辅助放射学与外科杂志，第12卷，第7期，第1089–1099页，2017年。[15] J. Bernal、J. S´anchez 和 F. Vilarino，“探索使用息肉外观模型进行自动息肉检测”，模式识别，第45卷，第9期，第3166–3182页，2012年。[16] M. Ganz、X. Yang 和 G. Slabaugh，“结肠镜窄带成像数据中的息肉自动分割”，IEEE生物医学工程汇刊，第59卷，第8期，第2144–2151页，2012年。[17] N. Tajbakhsh、S. R. Gurudu 和 J. Liang，“利用形状与上下文信息自动检测结肠镜视频中的息肉”，IEEE医学影像汇刊，第35卷，第2期，第630–644页，2015年。[18] P. Wang、X. Xiao、J. R. Glissen Brown、T. M. Berzin、M. Tu、F. Xiong、X. Hu、P. Liu、Y. Song、D. Zhang 等，“用于结肠镜检查期间息肉检测的深度学习算法的开发与验证”，自然·生物医学工程，第2卷，第10期，第741–748页，2018年。[19] Y. B. Guo 和 B. Matuszewski，“采用全卷积膨胀神经网络的GIANA息肉分割”，载于第14届计算机视觉、成像与计算机图形学理论及应用国际联合会议论文集。SCITEPRESS-科学与技术出版物，2019年，第632–641页。[20] D. Jha、P. H. Smedsrud、M. A. Riegler、D. Johansen、T. De Lange、P. Halvorsen 和 H. D. Johansen，“Resunet++：一种先进的医学图像分割架构”，载于2019年IEEE多媒体国际研讨会（ISM）。IEEE，2019年，第225–2255页。[21] R. Wang、S. Chen、C. Ji、J. Fan 和 Y. Li，“用于医学图像分割的边界感知上下文神经网络”，医学图像分析，第78卷，第102395页，2022年。[22] D. Jha、M. A. Riegler、D. Johansen、P. Halvorsen 和 H. D. Johansen，“DoubleU-Net：用于医学图像分割的深度卷积神经网络”，载于2020年IEEE第33届计算机医学系统国际研讨会（CBMS）。IEEE，2020年，第558–564页。[23] C.-H. Huang、H.-Y. Wu 和 Y.-L. Lin，“HardNet-MSEG：一种简单的编码器-解码器息肉分割神经网络，平均Dice系数超过0.9，帧率达86 fps”，arXiv预印本 arXiv:2101.07172，2021年。[24] P. Chao、C.-Y. Kao、Y.-S. Ruan、C.-H. Huang 和 Y.-L. Lin，“HarDNet：一种低内存流量网络”，载于IEEE/CVF国际计算机视觉会议论文集，2019年，第3552–3561页。[25] N. K. Tomar、D. Jha、M. A. Riegler、H. D. Johansen、D. Johansen、J. Rittscher、P. Halvorsen 和 S. Ali，“FANet：一种用于改进生物医学图像分割的反馈注意力网络”，IEEE神经网络与学习系统汇刊，2022年。[26] R. Garg、V. K. Bg、G. Carneiro 和 I. Reid，“单视图深度估计的无监督卷积神经网络：几何学来助力”，载于欧洲计算机视觉会议。施普林格，2016年，第740–756页。[27] C. Godard、O. Mac Aodha 和 G. J. Brostow，“具有左右一致性的无监督单目深度估计”，载于IEEE计算机视觉与模式识别会议论文集，2017年，第270–279页。[28] M. Jaderberg、K. Simonyan、A. Zisserman 等，“空间变换网络”，神经信息处理系统进展，第28卷，2015年。[29] T. Zhou、M. Brown、N. Snavely 和 D. G. Lowe，“从视频中无监督学习深度和自运动”，载于IEEE计算机视觉与模式识别会议论文集，2017年，第1851–1858页。[30] R. Mahjourian、M. Wicke 和 A. Angelova，“利用三维几何约束从单目视频中无监督学习深度和自运动”，载于IEEE计算机视觉与模式识别会议论文集，2018年，第5667–5675页。[31] Z. Yin 和 J. Shi，“Geonet：稠密深度、光流和相机位姿的无监督学习”，载于 IEEE计算机视觉与模式识别会议论文集，2018年，第1983–1992页。 [32] Z. Wang、A. C. Bovik、H. R. Sheikh 和 E. P. Simoncelli，“图像质量评估：从误差可见性到结构相似性”，IEEE图像处理汇刊，第13卷，第4期，第600–612页，2004年。 [33] L. Zhou、J. Ye、M. Abello、S. Wang 和 M. Kaess，“通过捆绑调整、超分辨率和截断损失进行单目深度估计的无监督学习”，arXiv预印本 arXiv:1812.03368，2018年。 [34] C. Wang、J. M. Buenaposada、R. Zhu 和 S. Lucey，“使用直接法从单目视频中学习深度”，载于IEEE计算机视觉与模式识别会议论文集，2018年，第2022–2030页。[35] C. Godard、O. Mac Aodha、M. Firman 和 G. J. Brostow，“深入探究自监督单目深度估计”，载于 IEEE/CVF国际计算机视觉会议论文集，2019年，第3828–3838页。 [36] X. Lyu、L. Liu、M. Wang、X. Kong、L. Liu、Y. Liu、X. Chen 和 Y. Yuan，“Hr-depth：高分辨率自监督单目深度估计”，载于 AAAI人工智能会议论文集，第35卷，第3期，2021年，第2294–2301页。 [37] S. Shao、Z. Pei、W. Chen、B. Zhang、X. Wu、D. Sun 和 D. Doermann，“微创手术场景中的单目深度估计自监督学习”，载于2021年 IEEE国际机器人与自动化会议（ICRA）。IEEE，2021年，第7159–7165页。 [38] S. Shao、Z. Pei、W. Chen、W. Zhu、X. Wu、D. Sun 和 B. Zhang，“内窥镜检查中的自监督单目深度与自运动估计：外观流来助力”，医学图像分析，第77卷，文章编号102338，2022年。[39] J.-W. Bian、H. Zhan、N. Wang、Z. Li、L. Zhang、C. Shen、M.-M. Cheng 和 I. Reid，“从视频中进行无监督尺度一致的深度学习”，国际计算机视觉杂志，第129卷，第9期，第2548–2564页，2021年。 [40] A. Kendall、Y. Gal 和 R. Cipolla，“利用不确定性对场景几何与语义任务的损失加权进行多任务学习”，载于 IEEE计算机视觉与模式识别会议论文集，2018年，第7482–7491页。 [41] V. Guizilini、R. Hou、J. Li、R. Ambrus 和 A. Gaidon，“用于自监督单目深度估计的语义引导表征学习”，载于ICLR，2020年。[42] H. Su、V. Jampani、D. Sun、O. Gallo、E. Learned-Miller 和 J. Kautz，“像素自适应卷积神经网络”，载于IEEE/CVF计算机视觉与模式识别会议论文集，2019年，第11 166–11 175页。 [43] A. F. Agarap，“使用整流线性单元（ReLU）的深度学习”，arXiv预印本 arXiv:1803.08375，2018年。 [44] K. He、X. Zhang、S. Ren 和 J. Sun，“用于图像识别的深度残差学习”，载于 IEEE计算机视觉与模式识别会议论文集，2016年，第770–778页。 [45] A. Paszke、S. Gross、S. Chintala、G. Chanan、E. Yang、Z. DeVito、Z. Lin、A. Desmaison、L. Antiga 和 A. Lerer，“PyTorch中的自动微分”，载于 NIPS-W，2017年。 [46] D. P. Kingma 和 J. Ba，“Adam：一种随机优化方法”，arXiv预印本 arXiv:1412.6980，2014年。[47] A. Rau 、 P. Edwards 、 O. F. Ahmad 、 P. Riordan 、 M. Janatka、L. B. Lovat 和 D. Stoyanov，“利用条件生成对抗网络进行隐式域适应，以实现内窥镜检查中的深度预测”，国际计算机辅助放射学与外科杂志，第14卷，第7期，第1167–1176页，2019年。 [48] H. Itoh、M. Oda、Y. Mori、M. Misawa、S.-E. Kudo、K. Imai、S. Ito、K. Hotta、H. Takabatake、M. Mori 等，“通过域转换并辅以保持朗伯反射特性的辅助任务进行无监督结肠镜深度估计”，国际计算机辅助放射学与外科杂志，第16卷，第6期，第989–1001页，2021年。 [49] S. Mathew、S. Nadeem、S. Kumari 和 A. Kaufman，“利用扩展方向性CycleGAN增强结肠镜检查中的有损图像转换”，载于 IEEE/CVF计算机视觉与模式识别会议论文集，2020年，第4696–4705页。 [50] J.-Y. Zhu、T. Park、P. Isola 和 A. A. Efros，“使用循环一致对抗网络进行非配对图像到图像转换”，载于 IEEE国际计算机视觉会议论文集，2017年，第2223–2232页。 [51] P. Azagra、C. Sostres、 ´A. Ferrandez、L. Riazuelo、C. Tomasini、O. L. Barbed、J. Morlana、D. Recasens、V. M. Batlle、J. J. G´omez-Rodr´ıguez、

等，“完整校准内窥镜检查流程的 Endomapper 数据集”，*arXiv* 预印本 *arXiv:2204.14240*，2022年。 [52] Y. Ma、X. Chen、K. Cheng、Y. Li 和 B. Sun，“Ldpolypvideo 基准：包含多种息肉的大规模结肠镜视频数据集”，载于医学图像计算与计算机辅助干预国际会议。施普林格，2021年，第387–396页。 [53] D. Eigen、C. Puhrsch 和 R. Fergus，“使用多尺度深度网络从单幅图像预测深度图”，神经信息处理系统进展，第27卷，2014年。 [54] V. Usenko、N. Demmel 和 D. Cremers，“双球面相机模型”，载于*2018*年*3D*视觉国际会议（*3DV*）。IEEE，2018年，第552–560页。 [55] L. Manfredi 和 A. Cuschieri，“用于无缆气动软体机器人的无线紧凑型控制单元（WICCU）”，载于*2019*年第二届 *IEEE*软体机器人国际会议（*RoboSoft*）。IEEE，2019年，第31–36页。