DOI:10.11784/tdxbz202310025

基于风格迁移的柔性输尿管内窥镜图像深度估计

辛运帏[1,2]，尹晶晶[1,2]，赵　煜[3]，代　煜[3]，崔　亮[4]，殷小涛[5]

（1. 南开大学网络空间安全学院，天津 300071；
2. 天津市网络与数据安全技术重点实验室，天津 300071；
3. 南开大学机器人与信息自动化研究所，天津 300350；4. 民航总医院泌尿外科，北京 100123；
5. 解放军总医院第四医学中心泌尿外科，北京 100048）

摘　要：输尿管内窥镜手术是目前针对肾结石的主流治疗方案，其外形细长、镜体柔软，能够灵活穿越人体自然腔道的内径狭窄的尿道和输尿管，检查视野范围更广，使医生能够更好地观察到病变区域. 但一般的输尿管内窥镜仅配备单目摄像头进行配合手术操作，无法借助额外设备获取数据导致了其图像信息的匮乏；同时，相比于胃肠、鼻镜等手术场景，本研究的肾内场景在不具备公开数据集的同时，图像质量参差不齐，表面纹理细节不足，孔洞区域褶皱少，受模糊反光等干扰大，都易使深度估计受到影响针对以上问题，提出了一种基于改进风格迁移模型的深度估计方法. 该方法首先根据术前 CT 图像重建肾脏内部腔道模型并提取中心路径，将虚拟内窥镜的摄像头设置在插值后的路径点上，构建了虚拟内窥镜漫游图像与深度估计图像一一对应的数据集，并基于此数据集训练了一个深度估计模型；随后，使用添加高效通道注意力（ECA）模块的改进风格迁移模型，将真实内窥镜图像域迁移至虚拟内窥镜图像域；最后，再将经由风格迁移产生的虚拟内窥镜图像送入上述训练得来的深度估计模型中，最终实现真实内窥镜图像的深度估计. 所提方法的可行性及有效性在输尿管钬激光碎石术的图像中得到验证.

关键词：深度估计；风格迁移；注意力机制；深度学习
中图分类号：TP391　　**文献标志码**：A　　**文章编号**：0493-2137（2025）01-0047-09

Depth Estimation for Flexible Ureteroscopic Images Based on Style-Transfer

Xin Yunwei[1,2]，Yin Jingjing[1,2]，Zhao Yu[3]，Dai Yu[3]，Cui Liang[4]，Yin Xiaotao[5]

（1. College of Cyber Science，Nankai University，Tianjin 300071，China；
2. Tianjin Key Laboratory of Network and Data Security Technology，Nankai University，Tianjin 300071，China；
3. Institute of Robotics & Automatic Information System，Nankai University，Tianjin 300350，China；
4. Department of Urology，Civil Aviation General Hospital，Beijing 100123，China；
5. Department of Urology，Fourth Medical Center of Chinese PLA General Hospital，Beijing 100048，China）

Abstract：Using flexible ureteroscopes is the mainstream treatment method for nephrolithiasis owing to their slender and flexible structure，allowing them to navigate through the narrow natural passages of the human body，such as the urethra and ureters. The elongated shape of these ureteroscopes provides a broad visual inspection range，enabling physicians to effectively observe affected areas. However，conventional ureteroscopes typically feature a single-camera system for surgical operations，leading to limited image data. This limitation results in insufficient information as additional imaging devices cannot be leveraged. Unlike surgical scenarios such as gastrointestinal or nasal endoscopy，the kidney's internal environment lacks publicly available datasets. The images captured during surgery exhibit varying quality，with insufficient surface texture details，fewer folds in cavity areas，and susceptibility to

收稿日期：2023-10-22；**修回日期**：2024-03-01.
作者简介：辛运帏（1965—　），女，教授，xinyw@nankai.edu.cn.
通信作者：代　煜，daiyu@nankai.edu.cn.
基金项目：国家重点研发计划资助项目（2017YFB1302800）.
Supported by the National Key Research and Development Program of China（No. 2017YFB1302800）.

interference such as blurriness and reflections. These challenges can significantly impact depth estimation. Herein, a depth estimation method is proposed, which involves leveraging an improved style-transfer model to address the aforementioned issues. The method begins by reconstructing the internal cavity model of the kidney based on preoperative computed tomography images and extracting the central path. The involved virtual endoscope's camera is then positioned at interpolated path points, creating a dataset that correlates virtual endoscope roaming images with depth estimation images. A depth estimation model is trained using this dataset. Subsequently, an improved style-transfer model incorporating an efficient channel attention module is employed to transfer the real endoscopic image domain to the virtual endoscopic image domain. Finally, the virtual endoscopic images generated through style-transfer are input into the previously trained depth estimation model, achieving depth estimation of real endoscopic images. The feasibility and effectiveness of the proposed method are validated using images obtained from ureteroscopic holmium laser lithotripsy procedures.

Keywords: depth estimation; style-transfer; attention mechanism; deep learning

泌尿系统结石是泌尿系统中常见的疾病之一,其发病率呈上升趋势[1]. 泌尿系统任何部位均可产生结石,而肾脏是最为常见的始发源,输尿管内窥镜是目前肾结石治疗的重要方式. 输尿管内窥镜的外形细长、镜体柔软,能够灵活穿越生理弯曲的尿道和输尿管,且其检查视野范围更广,使医生能够更好地观察到病变区域,减少了对患者的创伤,术后恢复较快. 然而人体自然腔道的内径狭窄[2-3],为了减少患者的痛苦,一般的输尿管内窥镜仅配备单目摄像头进行配合手术操作,容易导致其末端的视野范围受限、深度信息丢失,医生仅依靠经验对二维的影像进行三维的场景感知,操作难度较大.

图像深度是构建肾腔的手术场景、解决内窥镜视野受限问题的基础,对真实内窥镜图像深度信息的恢复,可以更好地帮助外科医生了解场景空间. 深度估计是计算机视觉领域的重要课题,目前基于深度估计的研究主要分为 3 类:基于设备获取的方法、基于传统几何的方法和基于深度学习的方法.

基于设备获取的方法是指基于 RGB-D 相机直接获取深度图像,此类相机使用激光雷达等辅助设备获取深度图[4],但由于肾腔内部狭窄,对硬件的尺寸要求严格,安装额外设备的可行性低. 同时,多设备的信号容易在手术环境中与其他金属器械相互影响,稳定性无法保证,所以不适用于本文输尿管软镜的应用场景.

基于传统几何的方法指的是通过计算机视觉几何理论方法,恢复单目摄像头拍摄图像的深度,主要分为从阴影中恢复形状法(shape from shading, SFS)[5]和从运动中恢复结构法(structure from motion, SFM)[6]. 从运动中恢复结构法是较为常用的方法,主要思想是基于多视觉几何原理,通过单目相机运动产生的不同视点处图像的视差来获取相应点的空间位置. 张建勋等[7]提出了一种基于 SFM 的肾腔孔洞重建方法,但该方法对增强图像的纹理特征要求高,给孔洞点云的提取及匹配造成了较大的困难.

基于深度学习的方法可以分为自监督、半监督和有监督. 自监督或半监督[8]的方法利用位姿估计网络或者光流估计网络来估计相邻图像帧的运动变化,从而估计出深度信息. 在自然图像领域中,Zhou 等[9]提出了一种同时训练单目深度和相机位姿的网络模型,通过求解目标视图与合成视图的像素差得到深度估计结果,内窥镜场景下不具备可行性. Ozyoruk 等[10]提出了一种无监督的方法,该网络模型可以同时估计猪胃肠内窥镜图像下的深度和位姿,但这类方法的估计效果非常依赖于器官表面图像的低反射度和丰富纹理,并且联合估计网络的不确定性大,所以不适用于输尿管内窥镜场景下的深度估计.

由于单目内窥镜数据本身没有深度标签,有监督的深度估计方法一般需要通过借助其他模态的术前图像进行辅助. 如 Visentini-Scarzanella 等[11]提出了一种结合术前 CT 数据的支气管镜深度估计方法,通过渲染得到的虚拟内窥镜图像,训练真实内窥镜与纹理无关的有监督估计网络模型. Mahmood 等[12]提出了一种基于实影渲染技术的深度估计方法,通过模拟光线进行 CT 重建模型的传播,基于合成的虚拟图像训练深度估计网络,再应用到真实内窥镜图像中.

因此对输尿管内窥镜真实图像进行深度估计存在以下难点. ①输尿管内窥镜的手术场景十分狭窄,无法借助额外设备获取数据导致了其图像信息的匮乏. 同时,相比于胃肠、鼻镜等手术场景,本研究的肾内场景在不具备公开数据集的同时,图像质量参差不齐,表面纹理细节不足,孔洞区域褶皱少,受模糊反光等干扰大,都易使深度估计受到影响. ②内窥镜手术的场景尺度很小,基于传统视觉的方法、基于深度学习中无监督和半监督的方法只能恢复出图像点的相对深度,由于没有真实场景下的尺度因子,为深度

估计带来许多困难. 基于深度学习的无监督方法中，由其他模态图像向真实图像进行迁移学习的方法是一种解决以上问题的可行思路.

本文提出的深度估计方法适用于输尿管内窥镜手术场景的特点，首先基于术前的 CT 影像重建得到肾腔模型并提取出其中心路径，将漫游相机位置设置在路径点上获取漫游-深度图像对数据集，并使用该数据集训练深度估计模型，再通过风格迁移的方法，将真实内窥镜图像迁移至虚拟内窥镜图像域，进而通过上述深度估计模型，最终完成深度估计. 所提的方法在肾内手术的内窥镜图像中表现很好.

1　网络结构

由于进行肾腔手术的真实内窥镜本身不具备深度信息，而基于虚拟漫游路径的图像可以计算真实的深度信息，因而可以通过图像风格上的变换，实现对真实内窥镜进行间接的深度估计. 所提方法的整体框架如图 1 所示，将真实内窥镜域的图像送入融合注意力机制的改进风格迁移模型中，将其迁移为虚拟内窥镜域的图像，再通过训练虚拟内窥镜漫游图像上的深度估计模型，完成间接的深度估计任务.

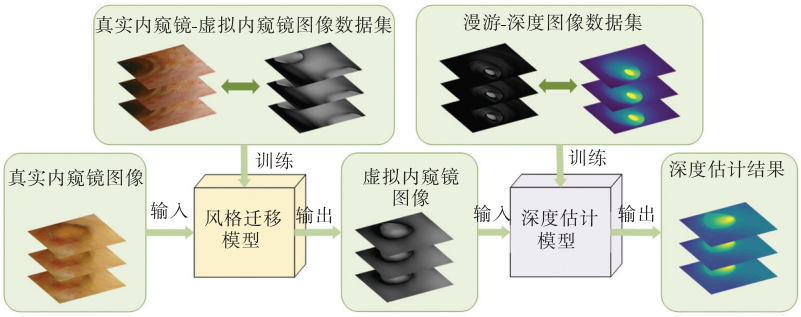


图 1　所提方法的总体结构

Fig.1　Overall framework of the proposed method

1.1　虚拟内窥镜图像数据集的构建

首先，对术前扫描得到的泌尿系统 CT 图像的肾脏区域进行分割，由于排泄期 CT 图像上造影剂充盈空腔，并且管道空腔的像素灰度值呈较为连续的分布，所以使用三维区域生长算法[13]对肾腔进行分割，选取的种子点向闭合连通邻域不断扩增，得到灰度值差异化的分割结果. 然后，选用实时交互性更好的面绘制算法中最为常用的移动立方体算法（marching cubes，MC）[14]，通过体素级的三角面片进行等值面的逼近及拼接，完成对管腔表面结构的三维重建，再经过网格平滑、漏洞修补和法向量调整，使其更加接近真实的多分支肾腔模型.

为了生成贴近真实手术内窥镜的漫游图像，提取出肾腔模型的中心路径点集，通过对模型三角剖分计算得到的维诺图作为基础，使用快速行进法求解维诺图上的程函方程，然后最小化可能路径的最大内切球的半径积分，其路径搜索的代价函数为

$$E_{\text{centerline}}(\boldsymbol{D})=\int_{\Omega}(E_{\text{energy}}(\boldsymbol{D}(s))+\omega)\mathrm{d}s \tag{1}$$

式中：E_{energy} 为曲线能量；ω 为规范项. 搜索中心路径点集旨在维诺图 Ω 中寻找一条累积能量 $E_{\text{centerline}}$ 最小的曲线 $\boldsymbol{D}(s)$，其初始点和终止点分别为 P_{start} 和 P_{end}. 规划出定步长的空间稠密路径点集 $\{P\}$，包含离散点坐标和管腔半径信息，使用非均匀有理 B 样条曲线进行插值平滑，最终规划得到管腔模型的中心路径，将虚拟相机设置在这些路径点上，进一步生成路径点上的漫游图像. 通过式（2）计算出其在当前相机视角下的真实深度值，并将虚拟内窥镜漫游图和深度图一一对应，生成漫游-深度图像对.

$$Z_{\text{real}}=\frac{2Z_{\text{far}}Z_{\text{near}}}{Z_{\text{far}}+Z_{\text{near}}-(Z_{\text{far}}+Z_{\text{near}})(2Z_{\text{buffer}}-1)} \tag{2}$$

式中：Z_{far} 和 Z_{near} 分别为相机聚焦时的剪切前平面和后平面的位置深度值；Z_{buffer} 为当前点的深度缓存值；Z_{real} 为当前像素点处的真实深度值.

1.2　虚拟内窥镜图像的深度估计

由于生成的虚拟内窥镜漫游图像均有一一对应的深度图像，所以采用有监督的方法进行深度估计，将其视作图像像素点的回归问题. 采用目前在自然图像领域的深度估计任务中较为常用的解码器-编码器（encoder-decoder）结构作为基础的网络框架构造出一个类似 U 型的网络模型，本文模型的设计思想基于 DenseDepth[15] 网络模型，使用在大型数据集 ImageNet[16] 上预训练的模型作为编码器，使用由卷积层和激活层组成的上采样模块作为解码器，并在编码器和解码器之间通过跳跃连接模块增强解码器模块捕获多层次特征信息的能力，使得网络可以同时获取结构信息和细节信息.

如图 2 所示，模型的编码器部分是基于 ImageNet 预训练的 MobileNetv2 模型[17]，主体结构为线性瓶颈(linear bottleneck，LB)的倒残差结构(inverted residual，IR)，在适应输入尺寸信息的同时保证有用信息的高维嵌入，且整体编码器结构相对轻量化. 模型的解码器结构基于编码器对称构造，由卷积层和激活层组成.

采用 3 种损失函数相结合的方法来提高模型训练的精度，即

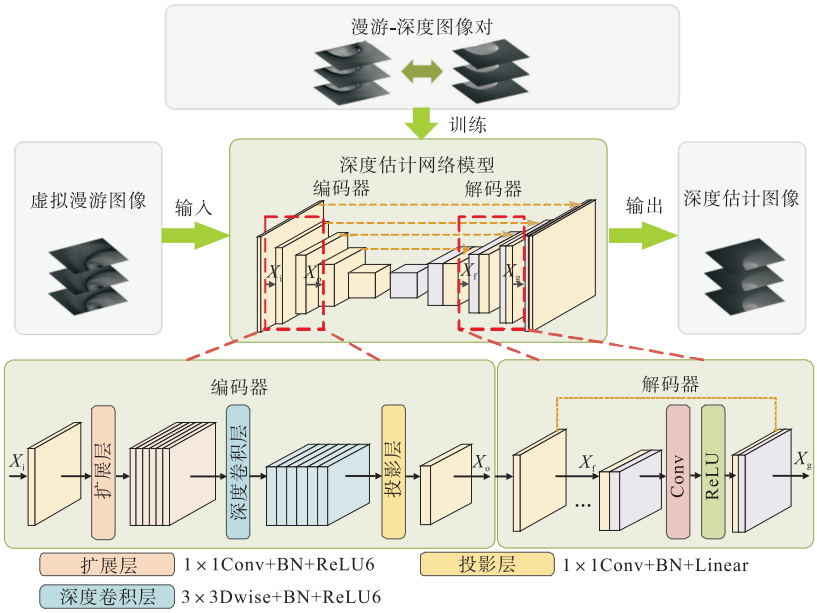


图 2 虚拟内窥镜图像的深度估计网络

Fig.2 Depth estimation network for virtual endoscopic images

$$L(y,\hat{y})=\lambda L_{\mathrm{MAE}}(y,\hat{y})+L_{\mathrm{GRAD}}(y,\hat{y})+L_{\mathrm{SSIM}}(y,\hat{y}) \tag{3}$$

式中：y 为图像的真实深度值；$\hat{y}$ 为预测深度值；L_{MAE} 为绝对误差；L_{GRAD} 为梯度误差；L_{SSIM} 为结构相似性损失函数；λ 设置为 0.1.

像素点处深度值较大时，计算该损失会过大，所以在送入模型训练之前，使用场景中的最大深度值对图像深度进行归一化处理.

1.3 基于风格转移的真实内窥镜图像深度估计

由于进行肾腔手术的真实内窥镜本身不具备深度信息，而基于虚拟漫游路径的图像可以计算真实的深度信息，所以可以通过图像风格上的变换，对真实内窥镜进行间接的深度估计. 但真实内窥镜图像存在出血红斑、气泡飞沫以及快速行进时造成的模糊等问题，与虚拟内窥镜的图像纹理信息相差较大，无法通过直接渲染的方法实现二者风格上的相互转换，所以考虑使用深度学习中风格迁移的方法将真实内窥镜图像转换为虚拟内窥镜图像，进而通过基于漫游-深度数据集训练完成的深度估计网络模型进行真实内窥镜的深度估计. 在常用于风格迁移任务的生成对抗网络中，循环生成对抗网络(cycle generative adversarial network，CycleGAN)[18]是无监督方法，打破了有监督方法对源域和目标域图像必须一一对应的条件，在两个域的图像并无匹配关系时，也可以通过训练得到较好的迁移效果. 由于本文所使用的虚拟内窥镜与真实内窥镜并不存在完全的一一对应的关系，所以选用该模型进行训练.

如图 3 所示，在 CycleGAN 网络模型中，A 表示处于真实内窥镜图像域的数据，B 表示处于虚拟内窥镜图像域的数据，模型最终需要实现 A 域和 B 域数据的互相风格迁移变换. 该模型由上下两个 GAN 网络组成，每个网络包含两个生成器 G、F 和两个判别器 D_A、D_B. 其中网络 1 需要实现从 A 域图像 Real_A 生成 B 域图像 Fake_B 再生成 A 域图像 Recover_A 这样一个风格迁移变换循环，此时通过计算初始 Real_A 图像和 Recover_A 图像的循环损失 Loss_cycle(A)以保证内容的相似性. 对于判别器而言，由于生成器的输出会随着迭代次数的增加更加贴近相应模态域的图像风格，所以它需要检测出生成器的图像为假，此时的对抗损失 Loss_gan(A)即在保证生成器和判别器的相互进化. 网络 2 与网络 1 过程相反，其循环损失为 Loss_cycle(B)，对抗损失 Loss_gan(B).

在训练中更新生成器 F 参数的时候，可能会出现 Fake_B 图像在风格和内容上与 Real_A 图像大相

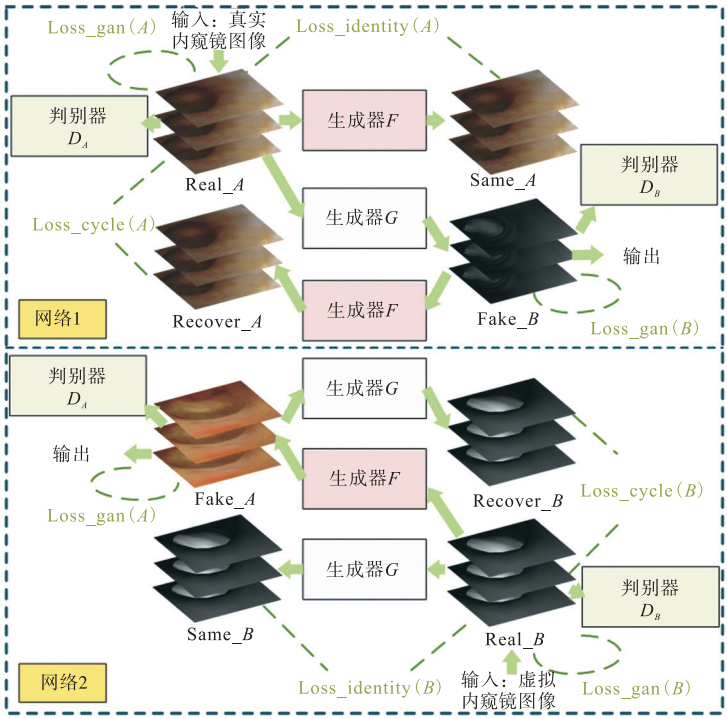


图 3　CycleGAN 网络模型
Fig.3　CycleGAN network model

径庭的情况，但由于 Loss_cycle(*A*)的存在，生成器 *F* 仍会随着迭代次数增加使得 Recover_*A* 尽力贴近原图像，所以需要将输入的 Real_*A* 图像经由生成器 *F* 生成 Same_*A* 图像，并通过计算二者之间的映射损失 Loss_identity(*A*)，保证图像在经由生成器进行风格迁移训练的同时，其图像内容不发生改变. CycleGAN 模型的总体损失函数为

$$\text{Loss}=\text{Loss\_gan}+\text{Loss\_cycle}+\text{Loss\_identity} \tag{4}$$

式中循环损失 Loss_cycle 、对抗损失 Loss_gan 及映射损失 Loss_identity 计算公式分别为

$$\begin{aligned}\text{Loss\_cycle}=&E_{A\sim P_{\text{data}}(A)}\left[\left\|F(G(A))-A\right\|_1\right]+\\&E_{B\sim P_{\text{data}}(B)}\left[\left\|G(F(B))-B\right\|_1\right]\end{aligned} \tag{5}$$

$$\begin{aligned}\text{Loss\_gan}=&\text{Loss\_gan}(A)+\text{Loss\_gan}(B)=\\&E_{A\sim P_{\text{data}}(A)}[\text{lb}D_A(A)]+\\&E_{B\sim P_{\text{data}}(B)}[\text{lb}(1-D_A(F(B)))]+\\&E_{B\sim P_{\text{data}}(B)}[\text{lb}D_B(B)]+\\&E_{A\sim P_{\text{data}}(A)}[\text{lb}(1-D_B(G(A)))]\end{aligned} \tag{6}$$

$$\begin{aligned}\text{Loss\_identity}=&E_{A\sim P_{\text{data}}(A)}\left[\left\|F(A)-A\right\|_1\right]+\\&E_{B\sim P_{\text{data}}(B)}\left[\left\|G(B)-B\right\|_1\right]\end{aligned} \tag{7}$$

式中：*G* 、*F* 分别为由 *A* 域迁移至 *B* 域和 *B* 域迁移至 *A* 域的生成器；D_A 、D_B 分别为 *A* 域图像和 *B* 域图像的判别器；$\|\cdot\|_1$ 表示 L1 范数.

在该模型的实验中，判别器使用块级对抗生成网络（patch-based generative adversarial network，PatchGAN）[19]模型，其通过对分块图像进行真假预测实现二分类判断. 生成器使用在经典 ResNet-50 模型中引入高效通道注意力（efficient channel attention，ECA）[20]模块的 ECA-Net 模型，以更好地提升网络模型在训练过程中对图像特征信息的学习能力. ECA 模块在直接对输入特征图 X_{i} 进行全局平均池化后，通过自适应计算与通道维度成比例的核大小 k，确定出通道交互的覆盖范围，从而基于快速 1×1 卷积层实现局部跨通道的注意力交互，最后再基于归一化的权重生成加权后的特征图 X_{o}. 该模块可以较好地在保证模块轻量化的基础上，实现有效的图像特征捕获，判别出生成器的图像为假，如图 4 所示.

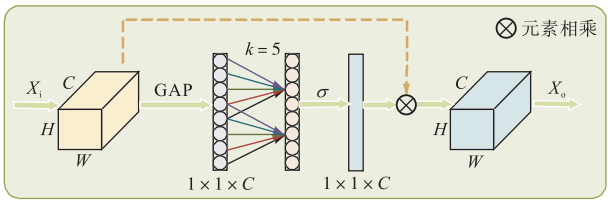


图 4　生成器中的 ECA 注意力模块
Fig.4　ECA attention module in the generator

图 4 中的 *H* 、*W* 、*C* 分别为特征图的高、宽和通道数，σ 为激活函数，局部交互核大小 k 与通道数 C 的关系为

$$k=\psi(C)=\left|\frac{\mathrm{lb}C}{\gamma}+\frac{b}{\gamma}\right|_{\mathrm{odd}} \tag{8}$$

式中：$|t|_{\mathrm{odd}}$ 表示距离 t 最近的奇数；γ 与 b 分别设置为 2 和 1.

在完成对真实内窥镜图像的风格迁移后，将输出的虚拟内窥镜域图像送入训练好的深度估计网络中，进而完成最终的深度估计任务，并进行实验结果分析.

2 实 验

2.1 虚拟内窥镜图像的深度估计结果

本章所有的模型实验，均是基于 CUDA 11.2 及 PyTorch 1.10 搭建的，使用型号为 NVIDIA GeForce RTX 3090 的 GPU 进行训练. 在深度估计模型实验中使用的是在 ImageNet 上预训练的 MobileNetv2 模型及其权重参数，设置训练轮次为 50，批大小为 4，学习率为 0.000 1，Adam 优化器参数分别为 0.9 和 0.999. 漫游-深度图像对中共有 15 000 个图像对，输入输出图像分辨率一致为 640 像素 × 480 像素，格式分别为 24 位彩色图及 24 位灰度图. 评估模型预测结果一般可以通过定性和定量方法实现，定性评估一般通过直观地观察模型输出的深度图，判断其与真实值的差异，定量评估一般通过准确度或误差等相关函数判断模型优劣，本文采取的指标分别为：δ_1、δ_2、δ_3、$L_{\mathrm{Abs\_rel}}$、$L_{\mathrm{Sq\_rel}}$、L_{RMSE}、$L_{\mathrm{RMSE(log)}}$ 和 L_{lg}.

为了更好地展示本文模型的预测结果，使用构建的漫游-深度图像对数据集，基于不同的损失函数进行实验，如表 1 所示. 其中，模型 1 设置损失函数为 MAE + SSIM，模型 2 设置为 RMAE + SSIM，模型 3 设置为 MAE + GRAD + SSIM. 可以看出，本文中采用的损失函数训练得到的结果在各个指标上均表现较优，验证了设置梯度损失及相似度损失的有效性.

表 1 虚拟内窥镜图像深度估计指标的计算结果

Tab.1 Calculation results of the depth estimation indexes for virtual endoscopic images

模型	δ_1	δ_2	δ_3	$L_{\mathrm{Abs\_rel}}$	$L_{\mathrm{Sq\_rel}}$	L_{RMSE}	$L_{\mathrm{RMSE(log)}}$	L_{lg}
模型 1	0.854	0.962	0.978	0.036	17.837	22.115	0.046	0.014
模型 2	0.913	0.987	0.995	0.022	13.475	22.008	0.051	0.017
模型 3	0.937	0.988	0.998	0.025	1.116	7.914	0.034	0.010

对本文深度估计方法进行定性评估，如图 5 所示，大部分图像都能较为准确地预测. 由于深度图本身是灰度图，不同灰度之间的差异不便直观观察，所以使用 viridis 颜色映射方法将灰度图转变为伪彩色图. 可以看出，模型在肾腔孔洞形态、位置不同时均能有较好的预测表现，对深度变化较大的孔洞内部的预测结果也与真实深度图深度信息基本一致. 但仅依靠灰度图不能非常准确地显示出预测值与真实值的差别，所以将二者的误差也进行了可视化展示，虚拟内窥镜的真实深度图本身即存在较为明显的等高线，所以大部分误差为边界等高线，其他区域误差较小，印证了定量指标的较好表现.

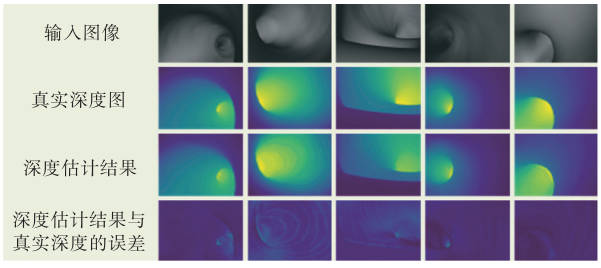


图 5 虚拟内窥镜深度估计结果

Fig.5 Results of virtual endoscopic depth estimation

2.2 基于风格迁移的深度估计结果

在风格迁移模型实验中，设置训练轮次为 200，批大小为 4，学习率为 0.000 2，Adam 优化器参数为 0.9 和 0.999. 训练数据为两个模态域. A 域为真实内窥镜图像，来自北京民航总医院对肾结石患者进行钬激光碎石取石手术的真实录像，手术使用奥林巴斯 URF-V2 输尿管电子内窥镜，其分辨率为 1 280 像素 × 720 像素，裁剪掉视频中的包含摄像头等无关信息得到 625 像素 × 479 像素的图像. B 域为虚拟内窥镜图像，是从深度估计模型的训练集中划分得到的部分图像. A、B 两个域的图像并不要求一一匹配. 分别采集 4 位患者术前 CT 图像数据和术中内窥镜视频数据，利用术前 CT 图像分割结果重建肾腔模型并生成 B 域虚拟内窥镜图像数据；对术中内窥镜视频数据截取其中有效数据帧作为 A 域真实内窥镜图像数据. 将两个模态域的图像数据按照 9∶1 的比例随机划分为训练集和测试集，最终划分数量如表 2 所示. 本研究仅收集临床数据和信息，为观察性研究，不会给受试者带来检查和治疗方面的任何风险.

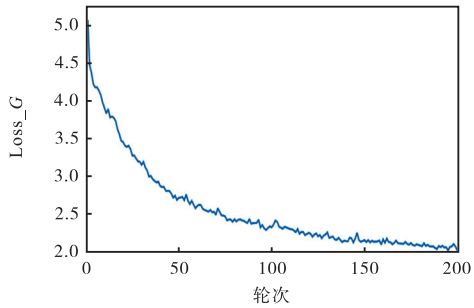
生成器损失 Loss_G 及判别器损失 Loss_D 的曲线如图 6 所示. 在经过 200 个轮次的训练之后，可以看到损失曲线趋于稳定，表明模型已经达到收敛点.

表 2　风格迁移图像数据划分数量

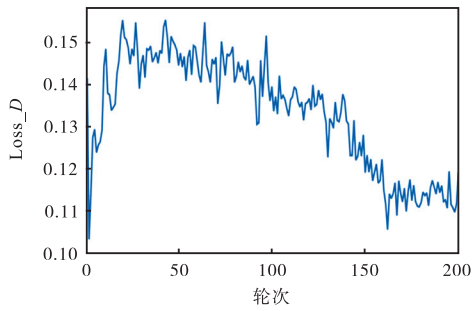
Tab.2　Number of style-transfer image data divisions

域	训练集图像数量	测试集图像数量
真实内窥镜 *A* 域	3 410	378
虚拟内窥镜 *B* 域	3 230	358

Loss_*G* 损失曲线在训练周期内稳步下降，在一个相对低的水平上波动，表明模型性能持续改善. Loss_*D* 的下降说明判别器逐渐难以区分生成的样本和真实样本，也就说明生成器的输出与真实样本之间的相似度变高.



（a）Loss_*G*

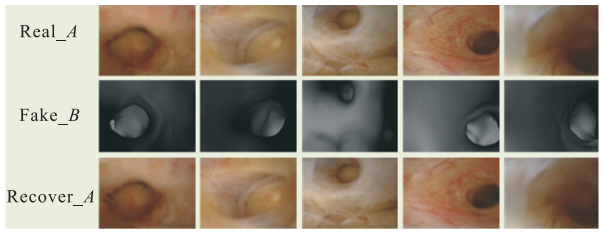


（b）Loss_*D*

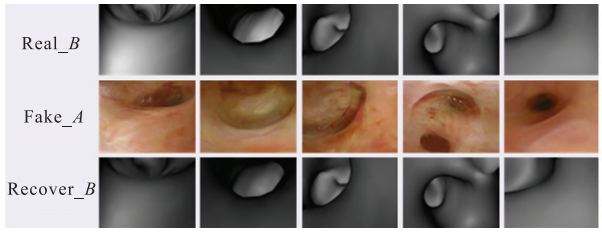
图 6　CycleGAN 模型的收敛曲线

Fig.6　Convergence curves of the CycleGAN model

对风格迁移后的两个模态域图像数据进行定性分析，将训练得到的 2 个生成器模型应用在测试集上，当输入为真实内窥镜 *A* 域的 Real_*A* 图像时，首先经过生成器 *G* 输出迁移为虚拟内窥镜 *B* 域的 Fake_*B* 图像，再经过生成器 *F* 输出迁移回到 *A* 域的 Recover_*A* 图像，同理，*B* 域的 Real_*B* 输入图像经由生成器 *F*、*G* 分别映射为 Fake_*A* 和 Recover_*B* 图像，如图 7 所示. 可以看出，经由转换的图像均能较好地在内容变化不大的前提下完成风格的变换，同时经由反向迁移时能够基本恢复为原输入图像. 由于单目内窥镜设备限制，并无配对的深度信息，所以需要专业领域内的医生来评估图像质量. 经过泌尿科两位医生的评估，一致认为本文方法相对最好，能够更好地关注到内窥镜图像的孔洞特征信息，预测出的边缘与真实内窥镜的孔洞位置基本相符，验证了本文方法的有效性. 两位医生分别来自民航总医院泌尿外科和解放军总医院第四医学中心泌尿外科.



（a）真实内窥镜 *A* 域迁移到虚拟内窥镜 *B* 域



（b）虚拟内窥镜 *B* 域反向迁移到真实内窥镜 *A* 域

图 7　真实内窥镜图像到虚拟内窥镜图像迁移及反向迁移结果

Fig.7　Results of style-transfer from the real endoscopic image to the virtual endoscopic image，and the reverse style-transfer results

从真实内窥镜图像风格迁移转换成虚拟图像会产生一定的误差，是客观存在的，虽然会有损失，经泌尿科两位医生的评估，本文方法在对比实验中表现最好，但由于此类风格迁移并没有通用的指标来定量地判定原始信息的损失程度，所以通过梯度幅度间接表示本文方法的效果. 视频帧图像的主要特征是肾腔孔洞，所以孔洞内外的深度梯度变化一般较快，对风格迁移之后的结果计算梯度，其结果如图 8 所示. 风格迁移结果的梯度幅度直观显示了孔洞位置的明显边界，表明网络对孔洞位置的信息损失不大；孔洞位置外的其余部分梯度连续，表示虚拟图像质量较好.

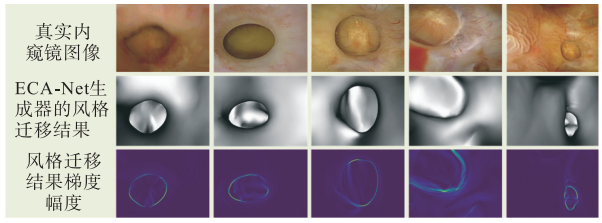


图 8　ECA-Net 风格迁移结果梯度幅度

Fig.8　Gradient magnitude of ECA-Net style transfer results

为了更好地判断风格迁移任务对真实内窥镜深度估计的影响，本文将真实内窥镜 *A* 域图像直接输

入训练好的深度估计模型，通过消融实验验证方法的有效性及必要性，如图 9 第 2 行所示，可以看出，经由风格迁移的图像能够较好地对几何特征及边缘特征明显的孔洞区域进行预测，这说明风格迁移应用于本文深度估计任务的可行性，但也可以直观观察到直接应用深度估计网络进行预测得到的结果在深度值上无法明确区分出亮暗、深浅特征的差异. 同时，本文使用基于深度学习 SfMLearner 模型[22]直接对单目真实内窥镜图像进行了深度估计，可以看出，该方法受内窥镜环境中的亮暗和纹理等变化的干扰较大.

最后，本文将基于 ResNet 模型与本文 ECA-Net 模型的实验结果进行了对比，如图 9 第 4、5 行和第 6、7 行所示，在图像整体亮度较高的情况下，模型的深度结果基本相近，但对于一些孔洞内部较暗的图像，添加 ECA 注意力机制后的训练模型能够更好地关注到内窥镜图像的孔洞特征信息，预测出的边缘与真实内窥镜的孔洞位置基本相符，验证了本文方法的有效性.

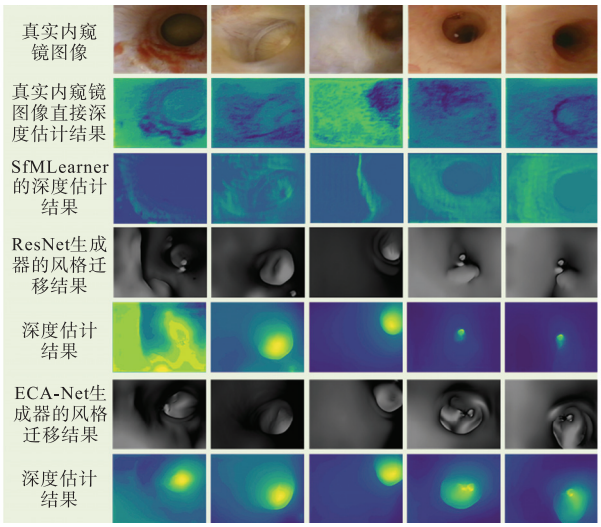


图 9 真实内窥镜图像深度估计结果

Fig.9 Results of real endoscopic image depth estimation

3 结 语

本文提出了一种针对真实的输尿管内窥镜手术图像的深度估计方法. 首先，通过患者术前 CT 影像的分割结果重建肾腔模型，由此获取其虚拟内窥镜漫游图像，进而构建漫游-深度图像数据集. 然后，训练基于编码器-解码器架构的深度估计模型，并创新性地使用基于 ECA 注意力机制的改进风格迁移模型，将真实内窥镜图像域的图像迁移至虚拟内窥镜图像域. 最后，将风格迁移后的虚拟内窥镜图像送入训练过的深度估计模型进行预测，最终完成深度估计任务. 实验分别从虚拟内窥镜图像的深度估计，以及真实内窥镜图像的风格迁移及深度估计验证了所提方法的准确性及有效性. 目前工作仍有许多方面需要改进，如：考虑虚拟内窥镜的相邻漫游图像和真实内窥镜的相邻视频帧的信息，引入其序列间相关性信息；尝试使用基于 Transformer 结构的特征提取器作为编码器.

参考文献：

[1] 李云鹏，吕建林. 人工智能技术在泌尿系结石中的应用与展望[J]. 临床泌尿外科杂志，2022，37(12)：957-959.
Li Yunpeng，Lü Jianlin. Application and prospect of artificial intelligence technology in urinary calculi[J]. Journal of Clinical Urology，2022，37(12)：957-959(in Chinese).

[2] 李建民，张增玉，赵建厂，等. 基于视觉反馈的输尿管软镜机器人运动补偿策略[J]. 天津大学学报(自然科学与工程技术版)，2021，54(7)：738-745.
Li Jianmin，Zhang Zengyu，Zhao Jianchang，et al. Motion compensation strategy for robot-assisted flexible ureteroscopy based on visual feedback[J]. Journal of Tianjin University(Science and Technology)，2021，54(7)：738-745(in Chinese).

[3] 洪 鹰，李辉鹭，肖聚亮，等. 柔性关节的复合动态面控制仿真研究[J]. 天津大学学报(自然科学与工程技术版)，2023，56(9)：973-984.
Hong Ying，Li Huilu，Xiao Juliang，et al. Simulation study of composite dynamic surface control of flexible-joint systems[J]. Journal of Tianjin University(Science and Technology)，2023，56(9)：973-984(in Chinese).

[4] 王太勇，孙浩文. 基于关键点特征融合的六自由度位姿估计方法[J]. 天津大学学报(自然科学与工程技术版)，2022，55(5)：543-551.
Wang Taiyong，Sun Haowen. Six degrees of freedom pose estimation based on keypoints feature fusion[J]. Journal of Tianjin University(Science and Technology)，2022，55(5)：543-551(in Chinese).

[5] Martino J M D，Qiu Q，Sapiro G. Rethinking shape from shading for spoofing detection[J]. IEEE Transactions on Image Processing，2020，30(11)：1086-1099.

[6] Song M，Watanabe H，Hara J. Robust 3D reconstruction with omni-directional camera based on structure from motion[C]//2018 IEEE International Workshop on Advanced Image Technology(IWAIT). Chiang Mai，

Thailand，2018：1-4.

［7］ 张建勋，韩明慧，代　煜. 面向低分辨率单目内窥镜图像的三维多孔结构重建[J]. 光学精密工程，2020，28(9)：2085-2095.
Zhang Jianxun，Han Minghui，Dai Yu. Three-dimensional porous structure reconstruction for low-resolution monocular endoscopic images[J]. Optics and Precision Engineering，2020，28(9)：2085-2095(in Chinese).

［8］ 姜　杉，张红运，杨志永，等. 基于无监督学习的三维肺部 CT 图像配准方法研究[J]. 天津大学学报(自然科学与工程技术版)，2022，55(3)：247-254.
Jiang Shan，Zhang Hongyun，Yang Zhiyong，et al. Research on a 3D lung computed tomography image registration method based on unsupervised learning[J]. Journal of Tianjin University(Science and Technology)，2022，55(3)：247-254(in Chinese).

［9］ Zhou T H，Brown M，Snavely N，et al. Unsupervised learning of depth and ego-motion from video[C]//2017 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Honolulu，USA，2017：1851-1858.

［10］ Ozyoruk K B，Gokceler G I，Bobrow T L，et al. EndoSLAM dataset and an unsupervised monocular visual odometry and depth estimation approach for endoscopic videos[J]. Medical Image Analysis，2021，71：102058.

［11］ Visentini-Scarzanella M，Sugiura T，Kaneko T，et al. Deep monocular 3D reconstruction for assisted navigation in bronchoscopy[J]. International Journal of Computer Assisted Radiology and Surgery，2017，12(7)：1089-1099.

［12］ Mahmood F，Chen R，Sudarsky S，et al. Deep learning with cinematic rendering：Fine-tuning deep neural networks using photorealistic medical images[J]. Physics in Medicine & Biology，2018，63(18)：185012.

［13］ 曹　彪. 基于区域生长的 OCT 图像分割算法研究[D]. 北京：北京理工大学，2015.
Cao Biao. Research of OCT Image Segmentation Algorithm Based on Region Growing Method[D]. Beijing：Beijing Institute of Technology，2015(in Chinese).

［14］ Okabe A，Boots B，Sugihara K，et al. Spatial Tessellations：Concepts and Applications of Voronoi Diagrams[M]. Hoboken：Wiley，2000.

［15］ Alhashim I，Wonka P. High quality monocular depth estimation via transfer learning[EB/OL]. https://arxiv.org/abs/1812.11941v2，2018-11-31.

［16］ Russakovsky O，Deng J，Su H，et al. Imagenet large scale visual recognition challenge[J]. International Journal of Computer Vision，2015，115(3)：211-252.

［17］ Sandler M，Howard A，Zhu M，et al. Mobilenetv2：Inverted residuals and linear bottlenecks[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City，USA，2018：4510-4520.

［18］ Zhu J Y，Park T，Isola P，et al. Unpaired image-to-image translation using cycle-consistent adversarial networks[C]// 2017 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Honolulu，USA，2017：2223-2232.

［19］ Isola P，Zhu J Y，Zhou T，et al. Image-to-image translation with conditional adversarial networks[C]// 2017 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Honolulu，USA，2017：1125-1134.

［20］ Wang Q，Wu B，Zhu P，et al. ECA-Net：Efficient channel attention for deep convolutional neural networks[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Las Vegas，USA，2020：11534-11542.

（责任编辑：王晓燕）