计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2024,33(1):58−67 [doi: 10.15888/j.cnki.csa.009343]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

结合改进注意力的肠胃镜图像深度估计①

林飞凡[1], 李　凌[2], 徐　强[2]

[1](湘潭大学 物理与光电工程学院, 湘潭 411105)
[2](苏州中科华影健康科技有限公司, 苏州 215123)
通信作者: 李　凌, E-mail: m13901540163@163.com

摘　要: 针对肠胃镜诊疗系统存在图像关键信息模糊和适应性差等问题, 提出一种结合改进注意力机制的循环生成对抗网络, 以实现对消化道深度信息的准确估计. 该网络以 CycleGAN 为基础网络, 结合双重注意力机制, 并引入残差门控机制和非局部模块来更全面地捕捉和理解输入数据的特征结构和全局相关性, 从而提高深度图像生成的质量和适应性; 同时采用双尺度特征融合网络作为判别器, 以提升其判别能力并平衡与生成器之间的工作性能. 实验结果表明, 在肠胃镜场景中预测效果良好, 相比其他无监督方法, 在胃道、小肠和结肠数据集上平均准确度分别提升了 7.39%、10.17% 和 10.27%. 同时, 在实验室人体胃道器官模型上也能够准确地估计出相对深度信息, 并提供精确的边界信息.

关键词: 肠胃镜图像; 深度估计; 生成对抗网络; 注意力机制; 双尺度特征

引用格式: 林飞凡,李凌,徐强.结合改进注意力的肠胃镜图像深度估计.计算机系统应用,2024,33(1):58–67. http://www.c-s-a.org.cn/1003-3254/9343.html

Depth Estimation of Gastrointestinal Endoscopy Images Using Improved Attention

LIN Fei-Fan[1], LI Ling[2], XU Qiang[2]

[1](School of Physics and Optoelectronics, Xiangtan University, Xiangtan 411105, China)
[2](Suzhou Ultimage Health Technology Co. Ltd., Suzhou 215123, China)

Abstract: In response to the key information blur in images and poor adaptability in the gastrointestinal endoscopy diagnosis and treatment system, this study proposes a cycle generative adversarial network (CycleGAN) combining an improved attention mechanism to accurately estimate the depth information of the digestive tract. Based on CycleGAN, the network combines a dual attention mechanism and introduces a residual gate mechanism and a non-local module to comprehensively capture and understand the feature structure and global correlation of input data, thereby improving the quality and adaptation of depth image generation. Meanwhile, a dual-scale feature fusion network is employed as the discriminator to improve the discrimination ability and balance the working performance between the generator and the discriminator. Experimental results show that the proposed method yields good prediction performance in the gastrointestinal endoscopy scenes. Its average accuracy of the stomach, small intestine, and colon datasets is improved by 7.39%, 10.17%, and 10.27% respectively compared with other unsupervised methods. Additionally, it can accurately estimate the relative depth information and provide accurate boundary information in the laboratory human gastric organ model.

Key words: gastrointestinal endoscopy image; depth estimation; generative adversarial network (GAN); attention mechanism; dual-scale feature

① 基金项目: 国家重点研发计划 (2020YFC2003802); 苏州市科技计划 (SYC2022109)
收稿时间: 2023-06-25; 修改时间: 2023-07-27; 采用时间: 2023-08-08; csa 在线出版时间: 2023-10-27
CNKI 网络首发时间: 2023-10-31

我国作为全球消化道系统疾病流行性最高的国家之一, 内窥镜检查成为国内最常见、最直接的检查手段. 近些年来, 微创手术逐渐和计算机技术相融合, 如外科医生结合手术经验和图像处理技术, 利用内窥镜导航系统[1]的深度估计技术, 可以在内窥镜手术[2]中为术者提供更准确的位置信息, 减少视觉误差. 在诊断方面, 可以利用深度信息来更好地检测病灶位置, 降低误诊概率.

单目深度估计作为场景感知的一部分, 从特定场景图像中产生对应像素级深度图, 同时也是三维重建[3]的关键技术. 相比于使用激光雷达[4]和双目定位[5]来获取深度信息, 单目视觉在内窥镜应用中有着更好的优势. 激光雷达成本过高、难以广泛应用. 双目定位难以匹配计算、计算量大, 也不利于肠胃镜微型化.

单目深度估计的方法一般分为有监督学习和无监督学习. 由于在单目内窥镜深度估计场景当中存在的局限性, 有监督学习需要真实深度图作为监督信号, 而医学数据和对应深度图难以获取并且代价昂贵, 近些年来大部分研究者都通过无监督学习来进行深度估计的研究[6]. Garg 等人[7]最早提出了使用无监督学习方法来进行深度估计, 该框架由估计深度的卷积神经网络 (convolutional neural network, CNN) 和图像重建模块构成. 利用图像重建模块重建出的新视图与输入图像之间的误差作为监督信号, 引导深度估计网络的训练. Zhou 等人[8]采用视频进行训练, 设计了一个深度估计网络 (DepthNet) 和一个姿态估计网络 (PoseNet), 分别估计深度和邻帧图像对的相机转换矩阵, 用转换矩阵来约束深度估计模型. Wang 等人[9]采用视觉里程计方法, 提出直接测距法 (direct visual odometry, DVO), 利用 3 帧图像之间的最小重建误差获得相机转换矩阵, 来得到准确的深度估计信息. Hur 等人[10]采用带有空间信息的场景流辅助估计深度, 帮助修正深度估计网络产生的误差. Cheng 等人[11]采用生成对抗网络 (generative adversarial network, GAN) 的方法, 提出了语义一致性迁移, 提取不同场景下相同语义信息对模型进行迁移引导, 从而估计深度信息.

上述方法在一般自然场景条件下已经相对成熟, 并能够取得良好的预测深度图. 然而与自然图像不同的是, 在肠胃镜图像中可能存在大量的噪声和冗余信息, 对图像的处理有较大的影响; 肠胃镜图像中存在更多的关键细节信息, 例如病变区域或者异常结构; 肠胃镜图像可能在不同的场景下采集, 例如不同的病人、不同的设备等, 对图像处理的适应性和泛化性具有较高的要求; 同时, 深度图的亮度、对比度、清晰度等特征, 也需符合医生的观察需求. 因此, 上述方法对肠胃镜图像缺乏适用性.

本文针对肠胃镜图像中存在的一系列问题, 采用循环生成对抗网络 (cycle generative adversarial network, CycleGAN) [12]为基础模型, 提出一种结合改进注意力机制的 CycleGAN 单目深度估计方法. 该方法大幅度降低了内窥镜图像的训练难度, 打开了肠胃镜深度图难以获取的局限性; 生成器网络结合双重注意力机制, 并在注意力机制中引入残差门控机制和非局部模块, 能够自动学习肠胃镜图像中的重要特征, 并将其加权融合到深度图像中, 从而提高生成图像的质量和真实度; 同时, 它还可以控制深度图像中的细节和风格, 使生成的图像更加符合预期的要求. 最后, 判别器采用一种双尺度特征融合网络来提高判别器的判别能力. 实验结果表明, 本文方法在不同猪胃肠道器官的内窥镜图像和人体胃道器官模型上均能取得较好的效果, 在大部分指标上都优于其他先进方法.

1 本文方法

本文提出的肠胃镜图像深度估计方法, 整体的网络结构流程如图 1 所示, 左右部分为一个对偶的正向与反向循环过程. 生成器 G 和 F 分别是 RGB 图像到深度图和深度图到 RGB 图像的映射, 判别器 D 对转换后的合成图像进行判别. 为了训练生成器和判别器的能力, 引入一个对抗性损失$Loss_{\mathrm{GAN}}$来提高它们彼此的能力. 同时, 为了确保生成的图像转换为输入图像时, 尽可能地与之相似, 在此引入一个循环一致性损失 $Loss_{\mathrm{cycle}}$, 来保证训练过程中最后的合成图像越来越像输入图像. 由于原始 CycleGAN 存在以下缺陷:生成器对噪声的处理能力较低; 生成器对具有较为复杂的细节信息的图像, 不能很好地提取和处理信息; 判别器判别能力低; 网络模型缺乏稳定性和泛化性. 生成的深度图整体准确度低, 并存在局部扭曲. 以问题为导向, 对网络结构中的生成器和判别器进行改进, 提高模型整体准确性, 稳定性和泛化性.

1.1 生成器网络结构

生成器的网络结构主要分为编码器、残差块和解

码器, 如图 2 所示. 编码器部分输入 256×256×3 的原始 RGB 图像, 首先进行图像增强, 将图像四边进行对称, 增大图像的分辨率, 再利用 3 个卷积网络进行特征提取. 第 1 个卷积网络采用 7×7 卷积核, 步幅为 1; 第 2、3 个卷积网络采用 3×3 卷积核, 步幅为 2. 解码器部分利用 2 个反卷积网络从高级特征中还原出低级特征, 每个反卷积网络采用 3×3 卷积核, 步幅为 2. 每个卷积和反卷积操作后均包含 1 个 InstanceNorm 归一化层和 1 个 ReLU 激活函数, 保证模型的稳定性. 最后得到 256×256×3 的深度图.

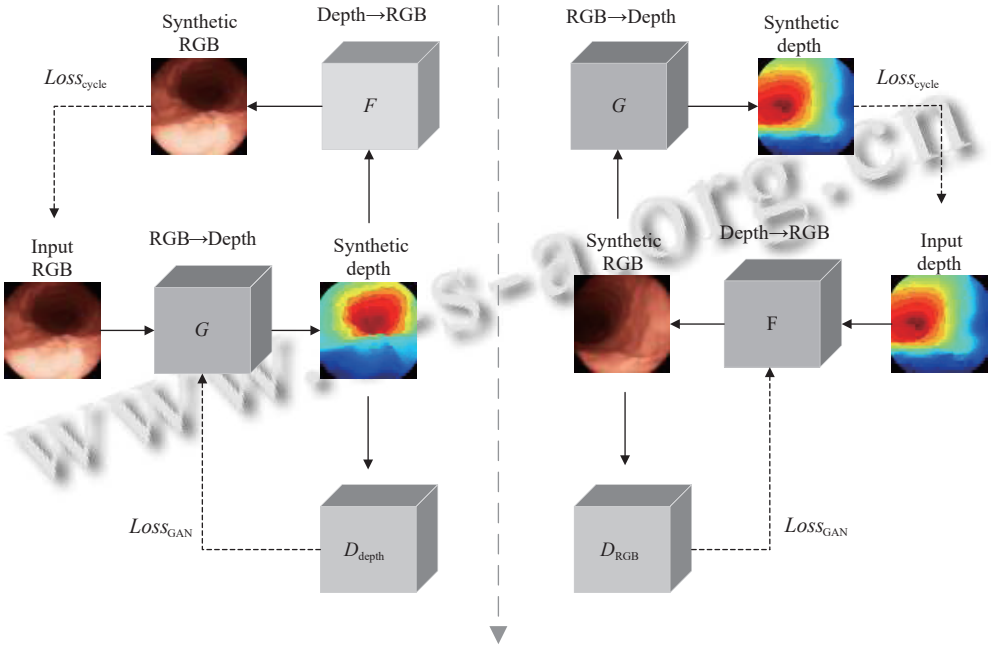

图 1 整体网络结构流程图

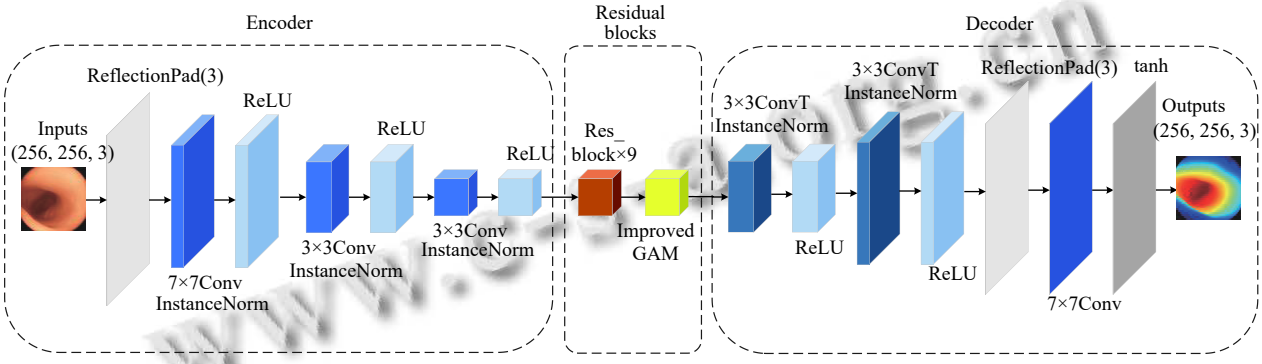

图 2 生成器网络结构

1.1.1 改进注意力机制

由于在肠胃镜图像深度估计任务当中, 由于其深度延展度小, 全局特征信息影响大, 传统的注意力机制在扩大感受野中丢失了边界轮廓和细节效果. 为了减少生成误差, 在生成器中结合全局注意力机制 (global attention mechanism, GAM)[13], 减少特征提取时全局信息的缺失, 放大全局通道和空间之间跨维度的交互作用, 使网络能够更好地关注重要的特征. 虽然生成过程中减少了信息缺失, 但对于肠胃镜图像中复杂的结构特征, 仍需提高生成图像的准确性和适应性. 因此本文方法使用非局部注意力机制和残差门控机制引入到 GAM 当中, 使得模型更加灵活、稳定和高效. 通过一个非局部块, 对输入数据的通道和空间信息进行加权, 以捕捉输入数据的全局相关性. 残差连接被用来连接原始输入和经过空间注意机制处理过的输出, 使得模型可以更容易地学习到残差信息. 同时, 门控机制被用

来控制残差连接的输出, 使得模型可以根据输入数据的不同自适应地调整输出数据的比例, 从而适应不同的场景. 改进的 GAM 过程如图 3 所示, 它是由通道注意力子模块、空间注意力子模块, 非局部块和残差门控子模块组合而成. 给定输入特征映射 $F_1 \in R^{C\times H\times W}$, 中间特征映射 F_2, F_3, F_4 和输出特征映射 F_5 定义为:

$$F_2 = M_C(F_1)\otimes F_1 \tag{1}$$

$$F_3 = M_S(F_{nl}(F_2))\otimes F_2 \tag{2}$$

$$F_4 = M_r(F_2)\oplus F_3 \tag{3}$$

$$F_5 = M_g(F_2)\otimes F_4 \tag{4}$$

其中, M_C 和 M_S 分别是通道注意力映射和空间注意力映射, M_{nl} 为非局部注意力映射, M_r 和 M_g 分别为残差映射和门控映射, ⊗表示元素相乘, ⊕表示元素相加. 通道注意力子模块主要分为 3 个操作 F_{sq}, F_{ex} 和 F_{sc}. F_{sq} 操作沿着空间维度进行特征压缩, 将每个二维特征通道变成一个实数, 输入 $W\times H\times C$ 特征图输出 $1\times 1\times C$ 特征图, 且具有全局感受野. F_{sq} 过程表示为:

$$z_c = F_{sq}(u_c) = \frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W}u_c(i,j) \tag{5}$$

其中, u_c 为输入特征矩阵第 c 个二维矩阵, c 表示通道. F_{ex} 过程表示为:

$$s_1 = F_{ex}(z,W) = \sigma(g(z,W)) = \sigma(W_2\delta(W_1 z)) \tag{6}$$

其中, z 为 F_{sq} 操作输出的结果. W_1 和 W_2 为两个全连接层操作, 利用全连接层来建立特征通道之间的相关性, 第 1 个全连接层来降低维度, 减少计算量, 第 2 个全连接层用于维度递增, 恢复到原来维度. 中间 δ 为 ReLU 函数, 来降低模型复杂度和提高训练能力. σ 为激活函数 Sigmoid, 最后得到权重 s_1. F_{sc} 使用 F_{ex} 输出的结果作为权重, 通过乘法逐通道加权到之前的特征上, 输入的 $1\times 1\times C$ 特征图还原到 $W\times H\times C$.

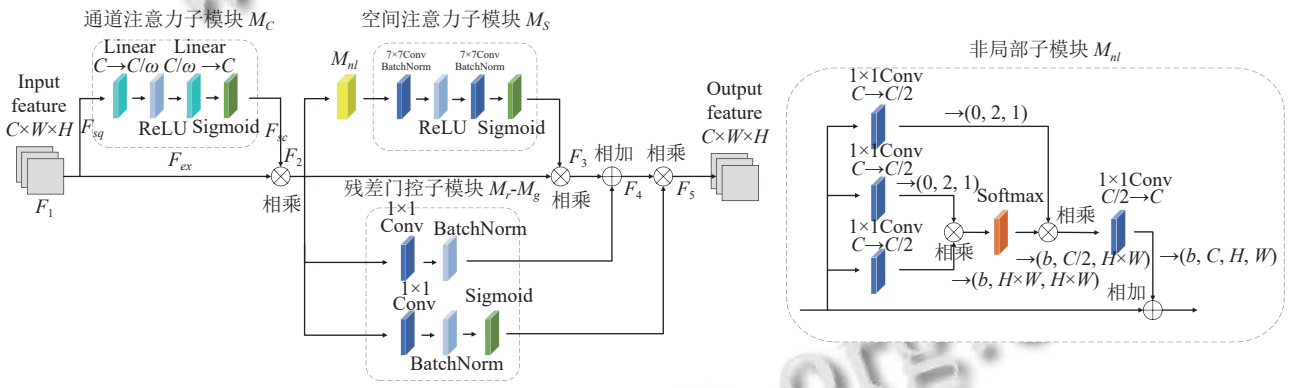

图 3 改进的 GAM 网络结构

在空间注意力子模块中, 通过一个非局部块和两个卷积网络进行空间信息融合. 非局部块该模块包含 3 个卷积层, 用于计算输入张量的全局关联性. 通过计算注意力权重和非局部响应来增强模型的感受野和建模能力, 提高生成图像的真实性和准确性. 在前向传播过程中, 该模块首先计算输入张量的全局关联性, 然后计算注意力权重张量和非局部响应张量, 最后将其与输入张量相加, 得到最终的输出张量. 为了减少信息的丢失, 相比其他传统注意力机制, 删除了池化层以保留最优特征信息. 具体过程表示为:

$$s_2 = \sigma(f^{7\times 7}(\delta(f^{7\times 7}(M_{nl}(F_2))))) \tag{7}$$

其中, $f^{7\times 7}$ 为 7×7 卷积核的卷积操作. 最后得到权重 s_2 与通道注意力子模块的输出特征 F_2 相乘, 输出特征尺寸保持不变.

残差门控子模块为一个残差连接和一个门控机制的组合, 作用于通道和空间注意力机制的输出特征, 利用门控机制来控制生成图像的特征, 使其更加符合输出需求. 通过将通道注意力机制处理过的输入和经过空间注意力机制的输出相加, 并使用 Sigmoid 函数来控制输出的比例, 增加模型的非线性能力和灵活性, 减少梯度消失和梯度爆炸的问题, 增加模型的泛化能力. 具体过程表示为:

$$F_5 = (f^{1\times 1}(F_2)\oplus F_3)\otimes(\sigma(f^{1\times 1}(F_2))) \tag{8}$$

1.1.2 改进残差块

在编码器和解码器的链接区域使用了 9 个重复的残差模块[14], 由编码器输出图像不同通道组合了图像

的不同特征, 根据这些特征来将图像的特征向量从源域转换到目标域. 针对特征图像的恢复能力和通道之间的相关性, 利用通道注意力子模块与残差模块相结合. 同时在通道注意力子模块中加入自适应平均池化层, 用于计算每个通道的重要性. 通过网络学习特征权重, 使有效的特征权重大, 无效或效果小的特征权重小, 进一步提升生成器的抗干扰能力和生成效果. 残差块网络结构如图 4 所示.

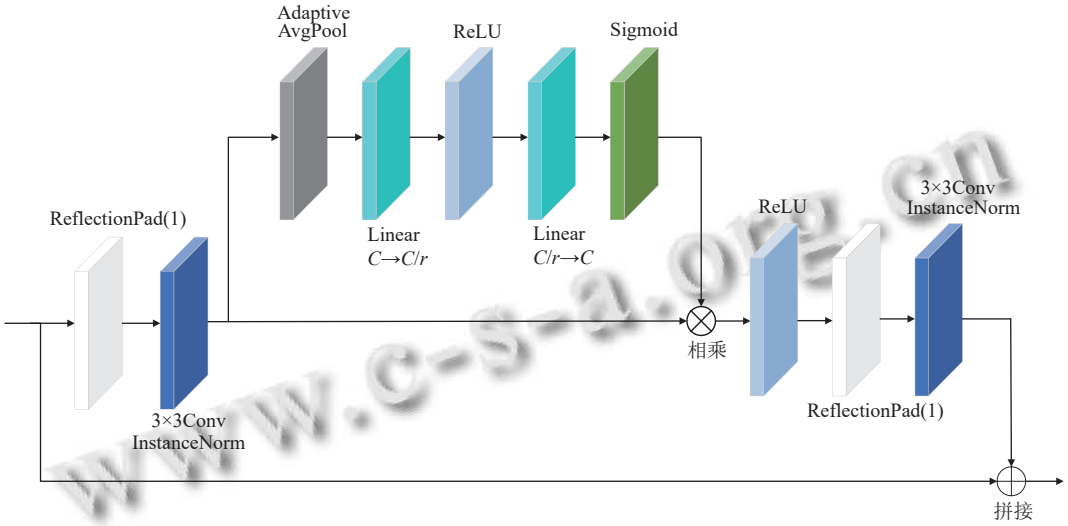

图 4 残差块结构

每块残差块中包含 2 个卷积网络和 1 个自适应通道注意力模块 (adaptive channel attention module, ACAM), 卷积核为 3×3, 步幅为 1. 将输入特征添加到输出特征中, 确保输入特征信息能够直接用作于后面的特征, 使得相应输入和输出不会有过大的偏差. 最后输出图像尺寸保持不变.

1.2 判别器网络结构

判别器网络对生成器生成的合成图像或者真实图像进行判别, 网络结构如图 5 所示. 针对判别器的判别能力低, 以及与生成器之间工作性能的平衡性, 本文采用一种双尺度特征融合的方法来进行优化. 判别器将输入图像经过两个不同尺度的分支网络进行特征提取, 这样可以更好地捕捉图像的全局和局部特征. 最后将两个不同尺度的特征进行融合判别, 以获取更全面的图像特征, 使判别器具有更好的鲁棒性和泛化能力, 从而提升判别器的判别能力.

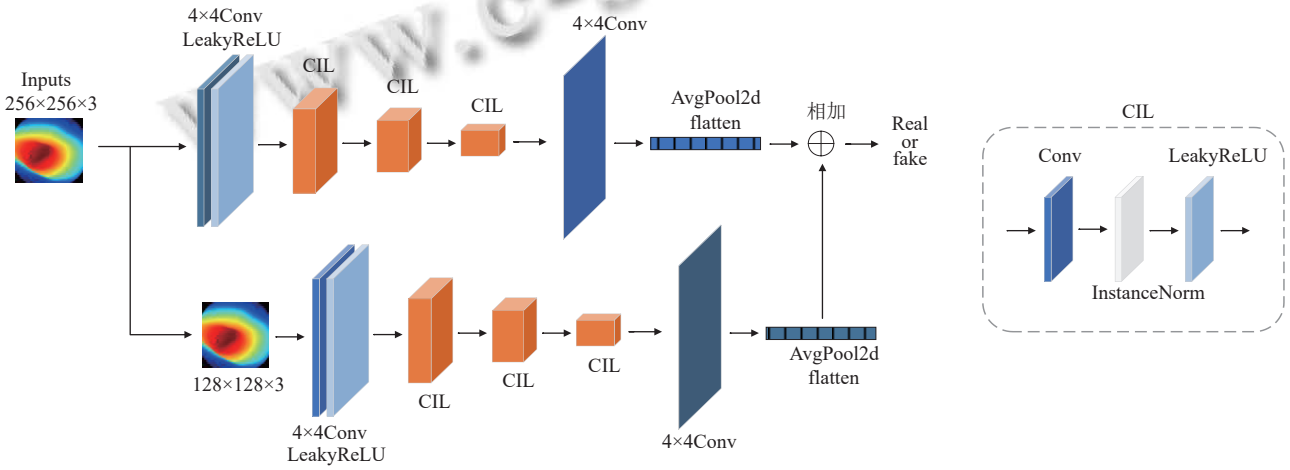

图 5 判别器网络结构

判别器输入 256×256×3 的 RGB 图像, 每个分支网络包含 4 个卷积网络来进行图像的特征提取, 每个卷积核均为 4×4, 步幅为 2. 归一化方式和生成器同样使用 InstanceNorm 层, 而防止判别器梯度稀疏, 采用 LeakyReLU 作为激活函数. 判别器为一个链式结构进行变换, 将最后一个卷积网络输出 30×30×1 的特征矩阵平展成一个一维矩阵进行判别.

为了帮助网络更快地收敛到最优解, 并且提高模型的性能, 在生成器和判别器当中使用均值为 0, 标准差为 0.02 的正态分布来初始化卷积层的权重, 并使用常数 0 来初始化卷积层的偏置.

1.3 损失函数

损失函数在网络模型训练当中起到至关重要的作用, 可以帮助 CycleGAN 生成高质量的图像, 并保持图像的一致性和准确性. CycleGAN 主要包含了对抗性损失和循环一致性损失, 即目标损失函数$Loss = Loss_{\mathrm{GAN}} + Loss_{\mathrm{cycle}}$. 本文方法采用 MSE 损失函数来测量对抗性损失$Loss_{\mathrm{GAN}}$, 它保证生成器和判别器相互进化, 进而保证生成器能产生更真实的图像. 它由两个映射对抗损失函数组成, 正向映射:

$$L_{\mathrm{GAN}}(G,D_Y,X,Y) = E_{y\sim p_{\mathrm{data}}(y)}[\log D_Y(y)] + E_{x\sim p_{\mathrm{data}}(x)}[\log(1-D_Y(G(x))] \quad (9)$$

反向映射:

$$L_{\mathrm{GAN}}(F,D_X,X,Y) = E_{x\sim p_{\mathrm{data}}(x)}[\log D_X(x)] + E_{y\sim p_{\mathrm{data}}(y)}[\log(1-D_X(G(y))] \quad (10)$$

MSE 损失函数可以帮助生成更加平滑和丰富的图像, 对于较小的像素级差异更加敏感.

循环一致性损失函数$Loss_{\mathrm{cycle}}$是指将生成的目标图像再次转换回原始图像, 并将其与原始图像进行比较. 这个损失函数可以帮助 CycleGAN 保持图像的一致性和准确性, 并避免出现不必要的变化. 本文利用 L1 损失函数测量循环一致性损失, L1 损失函数可以帮助生成更加清晰和锐利的图像, 损失函数设置为:

$$Loss_{\mathrm{cycle}} = E_{x\sim p_{\mathrm{data}}(x)}[\|F(G(x))-x\|_1] + E_{y\sim p_{\mathrm{data}}(y)}[\|G(F(y))-y\|_1] \quad (11)$$

2 实验过程

2.1 数据集

本文使用来自土耳其海峡大学生物医学工程研究所 (Institute of Biomedical Engineering, Bogazici University, Turkey) 提供的公开数据集[15], 其中包括离体猪消化道器官的标准胶囊内窥镜数据和合成的图像及其像素深度图. 为了获得准确的评价标准和泛化性, 本研究最终选用了合成的胃道, 小肠和结肠 3 组数据集, 分别包含 1 548 个、12 558 个和 21 887 个内窥镜图像, 分辨率为 320×320, 位深为 24 位, 其对应深度图位深为 32 位. 考虑到每组数据集的深度图风格略有不同和网络模型的泛化性, 实验随机选取了每组数据集中的 1 500 个图像进行分别实验, 并将它们深度图位深转换为 24 位. 最后将每组数据集分别随机划分 5 组, 其中 4 组作为训练集, 1 组作为测试集. 为了直观地展示预测结果, 将最后对预测结果转换为热力图[16].

2.2 环境配置与参数设置

本文实验在 PyTorch 框架中实现, 编码语言为 Python 3.6. 使用单块 NVIDIA GeForce RTX 3090 GPU, 操作系统为 Ubuntu 20.04, CUDA 版本为 11.3.

在网络训练阶段, batchsize 设为 2, 使用 Adam 优化器来训练网络, 参数$\beta_1 = 0.5$, $\beta_2 = 0.999$. 初始学习率设为 0.000 2, 共迭代 200 个 epoch. 其中, 前 100 个 epoch 学习率不变, 后 100 个 epoch 学习率线性下降到 0. 整个训练都在第 2.1 节的 3 组数据集上进行.

2.3 评价指标

使用单目图像深度估计中定量评价指标[17], 绝对相对误差 (*Abs_Rel*)、平均相对误差 (*Sq_Rel*)、均方根误差 (*RMSE*), 对数误差 (*LogRMS*) 和准确度 (δ). 其中误差值越小越好, 准确度越高越好. 指标具体公式分别如下:

$$Abs\_Rel = \frac{1}{N}\sum_{i=1}^{N}\frac{|D_i - D_i^*|}{D_i^*} \quad (12)$$

$$Sq\_Rel = \frac{1}{N}\sum_{i=1}^{N}\frac{|D_i - D_i^*|^2}{D_i^*} \quad (13)$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}|D_i - D_i^*|^2} \quad (14)$$

$$LogRMS = \sqrt{\frac{1}{N}\sum_{i=1}^{N}|\lg D_i - \lg D_i^*|^2} \quad (15)$$

$$\max\left(\frac{D_i}{D_i^*}, \frac{D_i^*}{D_i}\right) = \delta < T \quad (16)$$

其中, N 为像素总数, D_i为第 i 个像素的估计深度值,

D_i^*为第 i 个像素对应的真实深度值. 式 (16) 中, 阈值 $\delta < T = 1.25^{\theta}$, $\theta = 1,2,3$.

3 实验结果分析

3.1 定性分析

在本节中, 使用第 2.1 节的测试集进行深度图预测, 并与其他相关深度估计模型预测的深度图进行对比. 为了全面地分析本文方法的实验结果, 本研究选择了多个对比对象, 包括监督方法和无监督方法, 每个模型都使用训练结果最优的模型进行测试和定性分析. 部分深度估计结果如图 6 所示, 从左到右分别为胃道, 小肠和结肠的内窥镜图像.

通过对比实验结果发现, 在胃道数据集中, 本文提出的方法预测的深度图整体效果比其他方法更接近真实深度图. 然而, 在部分高光处理方面, 例如较近部分的褶皱受高光影响, 生成的深度图略显粗糙. Godard 等人[18]的方法受光影和纹理褶皱的影响较大, 不能很好地预测深度图, 其原因该网络对于具有大的深度变化的场景或者受场景结构的影响, 其深度估计精度会大幅度降低. 尽管 Isola 等人[19]的方法是监督学习, 但在该数据集上预测结果并不理想, 出现了严重的信息扭曲.

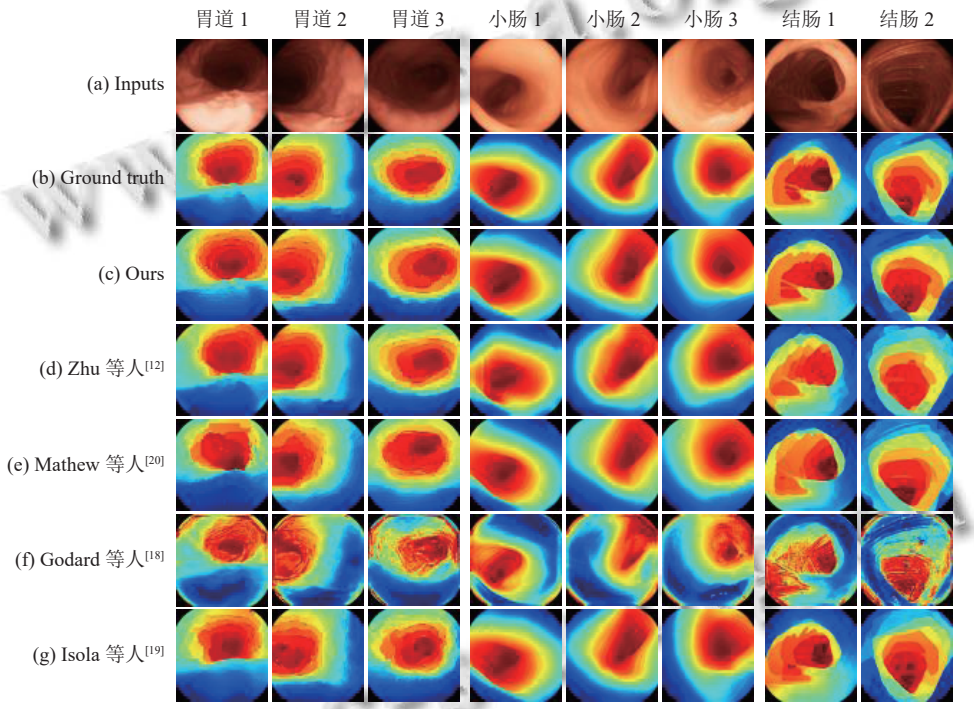

图 6 3 组数据集测试结果对比图

在小肠数据集上, 由于该数据集真实深度图更加光滑, 相对贴近其对应的原图, 因此训练难度小, 预测效果会更加明显. 本文方法在小肠数据集上的预测结果具有较好的效果, 整体效果和细节部分与真实深度图基本保持一致, 对边缘的预测具有准确的细节特征. Zhu 等人[12]的方法在预测结果的部分区域略有失真, 特别是在边缘信息和细节部分的处理上存在较大误差. Mathew 等人[20]的方法预测结果也具有较好的效果, 在整体效果上没有明显的误差, 但在部分细节处理上仍略逊于本文方法, 例如高光处的深度变化. 与胃道数据集上的问题相同, Godard 等人[18]的预测结果受到光影带来的影响较大.

结肠数据集的纹理褶皱较为明显, 且深度图的深度跨度较大, 相对于胃道和小肠数据集训练难度更高, 导致预测结果具有一定误差. 从图 6 中可以看出, 与其他方法相比, 本文方法的整体准确程度相对比较明显, 整体效果更加接近真实深度图. Zhu 等人[12]和 Mathew 等人[20]的方法在细节部分预测结果准确程度不高, 边缘信息误差较大, 缺乏对重要区域的关注. 而 Godard 等人[18]的方法预测的结果除了受光影的影响外, 同时还出现严重噪点. 相比于 Isola 等人[19]的监督方法, 本文方法在边缘信息和细节部分的预测更贴合真实深度,

比如近处结肠壁和洞深边缘. 本文所采用的方法在细节部分的把控具有良好的效果, 整体效果较好. 在上下肠胃镜图像中具有很好的泛化性和稳定性.

3.2 定量分析

除了直观的定性分析外, 还需要用准确的数据来体现预测结果的效果. 根据第 2.3 节的评估指标分别对 3 组数据集进行比较, 胃道数据集, 小肠数据集和结肠数据集对比结果分别如表 1–表 3 所示. 其中, “*”表示使用的是监督学习方法. 前 4 个评价指标参数越小越好, 后 3 个指标参数越大越好.

表 1 胃道数据集实验结果对比

方法	*Abs_Rel*	*Sq_Rel*	*RMSE*	*LogRMS*	δ_1	δ_2	δ_3
Zhu等人[12]	1.807	**0.991**	8.917	**0.351**	0.703	0.875	0.913
Mathew等人[20]	1.668	1.186	9.639	0.978	0.675	0.857	0.924
Godard等人[18]	2.487	1.551	10.086	0.618	0.407	0.637	0.779
Zhou等人[8]	2.461	1.989	10.034	0.706	0.506	0.757	0.859
Isola等人[19]*	1.824	1.124	9.569	0.643	**0.800**	0.920	0.964
Ours	**1.666**	1.023	**8.648**	0.422	0.778	**0.930**	**0.967**

表 2 小肠数据集实验结果对比

方法	*Abs_Rel*	*Sq_Rel*	*RMSE*	*LogRMS*	δ_1	δ_2	δ_3
Zhu等人[12]	3.752	1.312	9.277	1.156	0.695	0.865	0.927
Mathew等人[20]	2.107	1.037	9.357	4.335	0.698	0.869	0.918
Godard等人[18]	5.212	2.886	10.338	1.915	0.357	0.627	0.748
Zhou等人[8]	4.480	2.930	10.389	1.358	0.364	0.647	0.785
Isola等人[19]*	3.482	1.150	9.186	**0.802**	0.699	0.893	0.948
Ours	**1.472**	**0.868**	**8.293**	2.912	**0.823**	**0.944**	**0.973**

表 3 结肠数据集实验结果对比

方法	*Abs_Rel*	*Sq_Rel*	*RMSE*	*LogRMS*	δ_1	δ_2	δ_3
Zhu等人[12]	7.528	3.815	9.663	3.020	0.657	0.845	0.902
Mathew等人[20]	8.682	5.170	10.222	6.644	0.464	0.668	0.771
Godard等人[18]	7.686	3.493	10.109	**1.878**	0.389	0.611	0.744
Zhou等人[8]	8.157	4.030	10.236	2.997	0.473	0.682	0.800
Isola等人[19]*	**4.688**	1.300	8.084	6.370	0.769	0.858	0.887
Ours	4.843	**1.295**	**7.762**	3.135	**0.825**	**0.896**	**0.930**

在表 1 中看出, 本文提出的方法在胃道数据集中准确度明显高于其他无监督方法. 在 3 项准确度指标上比原网络的结果分别提升了 10.67%、6.29% 和 6.24%, 这是因为原网络缺乏对胃道中关键细节的学习, 同时受到无关信息的干扰. 但是本文方法在部分误差指标上表现不突出, 其原因是胃道数据集的真实深度图的整体相对比较粗糙, 细节部分不明显, 因此在提取更多细节时误差增大.

在表 2 和表 3 的数据集上, 本文所提出的方法在各项评估结果方面相较于其他方法有着显著的提升. 通过结合双注意力机制, 网络能够更好地捕捉图像的全局信息和局部信息, 从而更加有效地学习肠胃镜图像特征, 提高深度图像的质量. 同时, 结合双尺度判别器进行高性能判别, 使得网络模型在准确度指标方面有着更好的表现. 结合双注意力机制和双尺度判别器, 不仅提升了深度图像生成的质量和效率, 而且在处理肠胃边缘和细节部分时表现出更高的准确度. 前 3 项误差值与原网络相比均降低了 10.61% 以上, 在阈值 $\delta < 1.25$、1.25^2 和 1.25^3 方面, 准确度有着明显的提升.

因此, 在肠胃镜图像深度估计任务中, 本文方法明显优于其他无监督方法. 同时, 在与监督方法对比中也有一定的提升.

3.3 消融实验

为验证本文方法的有效性, 针对本文网络结构中结合的 GAM 以及残差块中的 ACAM 进行消融实验. 在 3 组数据集上进行了验证, 消融实验结果对比如图 7 所示.

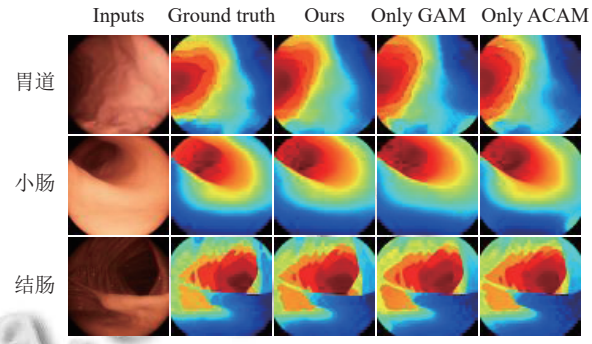

图 7 消融实验结果对比图

本文提出的循环生成对抗网络结合了改进的注意力机制, 其中双重注意力机制是本文方法的关键之一. 在生成器网络结构中, 本文方法相较于只保留 GAM 或 ACAM 的方法有着明显的优势. 当仅保留 GAM 时, 生成器失去了对于不同通道之间关系的感知能力, 导致生成的图像在纹理方面表现不够丰富, 从而在测试结果中纹理的效果有所下降. 而当仅保留 ACAM 时, 生成器失去了对于全局特征的感知能力, 无法整合特征图信息, 导致生成的图像在整体上表现不够自然、细节不够丰富. 在测试结果中可以看出, 缺失了局部细节的准确估计.

本文方法通过同时结合 GAM 和 ACAM, 能够兼顾不同通道之间的关系和全局特征的整合, 从而提高

了深度图像的质量和细节丰富度. 具体来说, ACAM 能够捕捉到不同通道之间的关系, 从而提高了图像的纹理丰富度; 而 GAM 能够整合特征图信息, 从而提高了深度图像的整体自然度和细节丰富度. 通过双重注意力机制的结合, 本文方法能够更全面地捕捉和理解输入数据的特征结构和全局相关性, 从而提高深度图像生成的质量和多样性.

表 4 为消融实验的评价指标结果, 与表 1–表 3 中本文方法的评价指标结果进行了对比. 结果表明, 本文方法在所有准确度指标上均优于仅采用单一注意力引导的网络. 进一步与原网络结果进行对比, 可以明显看出, 单注意力引导的网络在准确度指标上也得到了提升. 本文方法的优越性在于结合了双注意力机制和双尺度判别器, 能够更好地捕捉图像的全局信息和局部信息, 提高了图像生成的质量和效率. 因此, 在消融实验中, 本文方法表现出更好的性能, 为图像生成任务提供了更为可靠的解决方案.

表 4 消融实验评价指标结果

评价指标	Only GAM			Only ACAM			Ours		
	胃道	小肠	结肠	胃道	小肠	结肠	胃道	小肠	结肠
Abs_Rel	1.913	2.660	8.146	**1.648**	1.689	7.968	1.666	**1.472**	**4.843**
Sq_Rel	1.077	0.960	2.193	1.051	**0.770**	3.556	**1.023**	0.868	**1.295**
RMSE	9.054	8.601	9.422	8.924	**7.704**	8.566	**8.648**	8.293	**7.762**
LogRMS	0.659	**1.287**	**0.588**	0.528	3.363	0.798	**0.422**	2.912	3.135
δ_1	0.713	0.741	0.670	0.716	0.786	0.789	**0.778**	**0.823**	**0.825**
δ_2	0.885	0.885	0.810	0.881	0.889	0.882	**0.930**	**0.944**	**0.896**
δ_3	0.948	0.929	0.891	0.943	0.933	0.922	**0.967**	**0.973**	**0.930**

3.4 体模实验

为了进一步验证本文方法的泛化性, 本文使用实验室人体胃道器官模型进行测试. 该模型采用电子内窥镜进行胃道图像采集, 并将采集到的数据集使用训练好的胃道权重模型中进行测试. 实验中使用的模型为 IM8141-ERCP 训练模型, 图像采集设备为电子内窥镜, 型号 JIAWANG DWE-82L, 内窥镜图像处理器型号为 ShenDa EV-210A, 体模及设备如图 8 所示.

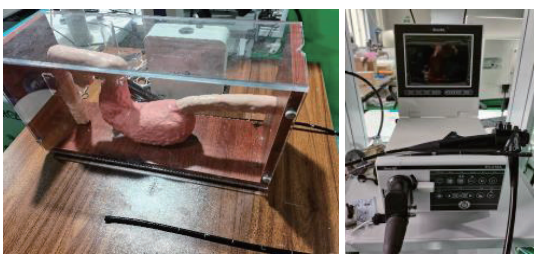

图 8 体模与内窥镜设备

部分测试结果如图 9 所示, 结果表明该模型能够较为准确地估计出相对深度信息, 并提供精确的边界信息. 此外, 深度连续, 且没有出现错误或扭曲的情况. 这些结果表明, 本文提出的方法具有高准确性、泛化性和实用价值, 能够有效地应用于实际的肠胃镜诊疗系统中.

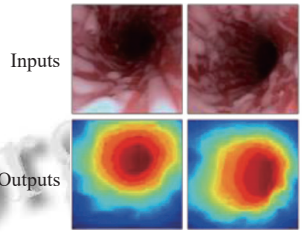

图 9 体模测试结果

需要注意的是, 实验室人体胃道器官模型与真实人体胃道存在一定的差异, 因此在实际应用中, 需要根据实际情况进行调整和优化. 此外, 本文方法还有一些局限性, 例如对于极端情况下的深度估计可能存在误差, 需要进一步改进和优化. 但是, 本文提出的方法为肠胃镜诊疗系统的深度信息估计提供了一种有效的解决方案, 具有重要的实际应用价值.

4 结语

针对消化道深度信息的准确估计, 本文提出的一种结合改进注意力机制的 CycleGAN 深度估计方法, 用于提升内窥镜对肠胃镜图像的适应性和关键信息的提取, 以及解决深度图边缘和细节信息模糊等问题. 实验结果表明, 结合改进注意力机制能够获得较为准确的深度估计结果. 相比其他相关深度估计模型, 本文方法预测的结果整体效果更接近真实深度图, 准确度更高, 对边缘信息和细节部分具有良好的预测能力. 此外, 实验室使用人体胃道器官模型进行测试, 网络模型能够得到准确的深度信息, 边缘信息也处理较好. 由此证明, 本文方法在肠胃镜图像应用中, 具有极佳的效果和较高的准确度. 在胃道、小肠, 结肠和体模测试中得出该网络模型具有优秀的泛化性和鲁棒性, 进一步推进了深度估计算法在医学领域中的实用化进展, 提升了肠胃镜的图像处理能力.

参考文献

1 崔曦雯, 陈芳, 韩博轩, 等. 虚拟内窥镜图像增强膝关节镜

手术导航系统. 中国生物医学工程学报, 2019, 38(5): 558–565.
2 胡天策, 蔡俊锋, 徐榕, 等. 基于内窥镜单目视觉手术导航的测距方法. 中国组织工程研究与临床康复, 2008, 12(22): 4241–4245.
3 Izadi S, Kim D, Hilliges O, *et al*. KinectFusion: Real-time 3D reconstruction and interaction using a moving depth camera. Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology. Santa Barbara: ACM Press, 2011. 559–568.
4 郑德华. ICP 算法及其在建筑物扫描点云数据配准中的应用. 测绘科学, 2007, 32(2): 31–32. [doi: 10.3771/j.issn.1009-2307.2007.02.009]
5 黄鹏程, 江剑宇, 杨波. 双目立体视觉的研究现状及进展. 光学仪器, 2018, 40(4): 81–86.
6 Zhao CQ, Sun QY, Zhang CZ, *et al*. Monocular depth estimation based on deep learning: An overview. Science China Technological Sciences, 2020, 63(9): 1612–1627. [doi: 10.1007/s11431-020-1582-8]
7 Garg R, Kumar BGVK, Carneiro G, *et al*. Unsupervised CNN for single view depth estimation: Geometry to the rescue. Proceedings of the 14th European Conference on Computer Vision. Amsterdam: Springer, 2016. 740–756.
8 Zhou TH, Brown M, Snavely N, *et al*. Unsupervised learning of depth and ego-motion from video. Proceeding of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 1851–1858.
9 Wang CY, Buenaposada JM, Zhu R, *et al*. Learning depth from monocular videos using direct methods. Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2018. 2022–2030.
10 Hur J, Roth S. Self-supervised monocular scene flow estimation. Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 7396–7405.
11 Cheng B, Saggu IS, Shah R, *et al*. S^3Net: Semantic-aware self-supervised depth estimation with monocular videos and synthetic data. Proceedings of the 2020 European Conference on Computer Vision. Glasgow: Springer, 2020. 52–69.
12 Zhu JY, Park T, Isola P, *et al*. Unpaired image-to-image translation using cycle-consistent adversarial networks. Proceeding of the 2017 IEEE International Conference on Computer Vision. Venice: IEEE, 2017. 2242–2251.
13 Liu YC, Shao ZR. Global attention mechanism: Retain information to enhance channel-spatial interactions. arXiv:2112.05561, 2021.
14 He KM, Zhang XY, Ren SQ, *et al*. Deep residual learning for image recognition. Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 770–7785.
15 Ozyoruk KB. Quantitative evaluation of endoscopic SLAM methods: EndoSLAM dataset. Proceedings of the 2020 IEEE International Conference on Robotics and Automation. 2020. 1–8.
16 Gonzalez RC, Woods RE, 著; 阮秋琦, 阮宇智, 译. 数字图像处理. 第 4 版, 北京: 电子工业出版社, 2020.
17 温静, 杨洁. 基于场景对象注意与深度图融合的深度估计. 计算机工程, 2023, 49(2): 222–230.
18 Godard C, Aodha OM, Firman M, *et al*. Digging into self-supervised monocular depth estimation. Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision. Seoul: IEEE, 2019. 3827–3837.
19 Isola P, Zhu JY, Zhou TH, *et al*. Image-to-image translation with conditional adversarial networks. Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 1125–1134.
20 Mathew S, Nadeem S, Kumari S, *et al*. Augmenting colonoscopy using extended and directional CycleGAN for lossy image translation. Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 1–10.

(校对责编: 孙君艳)