E-mail: jig@aircas.ac.cn
Website: www.cjig.cn
Tel: 010-58887035
中国图象图形学报
JOURNAL OF IMAGE AND GRAPHICS

中图法分类号:TP391.4　　**文献标识码**:A　　**文章编号**:1006-8961(2025)07-2542-16

论文引用格式:Liu M, Qin D X, Han Y B, Chen X and Wang Y N. 2025. Multiscale dynamic visual network for the scene segmentation of surgical robots. Journal of Image and Graphics, 30(7):2542-2557(刘敏, 秦敦璇, 韩雨斌, 陈祥, 王耀南. 2025. 多尺度动态视觉网络的手术机器人场景分割. 中国图象图形学报, 30(7):2542-2557)[DOI:10.11834/jig.240385]

多尺度动态视觉网络的手术机器人场景分割

刘敏[1,2,3],秦敦璇[1,2],韩雨斌[1,2,3*],陈祥[1,2],王耀南[1,2,3]

1. 湖南大学电气与信息工程学院,长沙　410082; 2. 机器人视觉感知与控制技术国家工程研究中心,长沙　410082;
3. 湖南省生物医学图像智能处理国际科技创新合作基地,长沙　410082

摘　要:**目的**　机器人辅助腹腔镜手术指的是临床医生借助腔镜手术机器人完成外科手术。然而,腔镜手术在密闭的人体腔道完成,且分割目标的特征复杂多变,对医生的手术技能有较高要求。为辅助医生完成腔镜手术,提出一种高精度的腔镜手术场景分割方法,并搭建分体式腔镜手术机器人对所提出的方法进行了验证。**方法**　首先,提出了多尺度动态视觉网络(multi-scale dynamic visual network, MDVNet)。该网络采用编码器—解码器结构。在编码器部分,动态大核卷积注意力模块(dynamic large kernel attention module, DLKA)可以通过多尺度大核注意力提取不同分割目标的多尺度特征,并动态选择机制进行自适应的特征融合。在解码器部分,低秩矩阵分解模块(low-rank matrix decomposition module, LMD)引导不同分辨率的特征图进行融合,可以有效滤除特征图中的噪声;边界引导模块(boundary guided module, BGM)可以引导模型学习手术场景的边界特征。最后,展示了基于Lap Game腹腔镜模拟器搭建的分体式腔镜手术机器人,网络模型的分割结果可以映射在手术机器人的视野中,辅助医生进行腔镜手术。**结果**　MDVNet在3个手术场景数据集上取得了最先进的结果,平均交并比分别为51.19%、71.28%和52.47%。**结论**　本文提出了适用于腔镜手术场景分割的多尺度动态视觉网络MDVNet,并在搭建的分体式腔镜手术机器人上对所提出方法进行了验证。代码开源地址为:https://github.com/YubinHan73/MDVNet。

关键词:腔镜手术机器人;语义分割;大核卷积;低秩矩阵分解(LMD);边界分割

Multiscale dynamic visual network for the scene segmentation of surgical robots

Liu Min[1,2,3], Qin Dunxuan[1,2], Han Yubin[1,2,3*], Chen Xiang[1,2], Wang Yaonan[1,2,3]

1. College of Electrical and Information Engineering, Hunan University, Changsha　410082, China;
2. National Engineering Research Center of Robot Visual Perception and Control Technology, Changsha　410082, China;
3. International Scientific and Technological Innovation Cooperation Base for Biomedical Image Processing, Changsha　410082, China

Abstract: **Objective** Robot-assisted endoscopic surgery is performed with the help of intelligent endoscopic surgical robots to effectively reduce trauma, shorten the recovery period, and improve surgical success rates. Endoscopic surgery scene segmentation means the use of deep learning techniques to accurately segment the entire surgical scene; the targets include anatomical areas and instruments. However, endoscopic surgery is completed in a closed human body cavity, and the whole process is accompanied by frequent cutting, traction, and other surgical operations, making the features of the seg-

收稿日期:2024-07-11;**修回日期**:2024-10-01;**预印本日期**:2024-10-08

*通信作者:韩雨斌　hyb_hnu@163.com

基金项目:国家重点研发计划资助(2022YFE0134700);国家自然科学基金项目(U22B2050,62221002,62425305);长沙市科技局重大科技项目(kq2102009)

Supported by: National Key R&D Program of China(2022YFE0134700); National Natural Science Foundation of China(U22B2050, 62221002, 62425305); Science and Technology Program of Changsha(kq2102009)

mentation targets complex and variable. In advancing robot-assisted surgery, developing high-precision surgical scene segmentation algorithms that can assist surgeons in performing surgeries is crucial. In this work, we propose an innovative surgical scene segmentation network named multiscale dynamic visual network (MDVNet), which aims to address three major challenges in endoscopic surgical scene segmentation: target size variation, intraoperative complex noises, and indistinguishable boundaries. **Method** MDVNet adopts an encoder-decoder structure. For the encoder, a dynamic large kernel convolutional attention (DLKA) module can extract multiscale features of the surgical scene is applied. The DLKA module consists of multiple branches each equipped with large kernel convolutions of different sizes, allowing the network to capture details and a wide range of features of targets with different sizes (7, 11, and 21 in this work). The dynamic selection mechanism also helps to adaptively fuse features to meet the needs of different sized segmentation targets in endoscopic surgery scene. This newly designed module directly addresses the problem of target size variability, which is a major obstacle faced by previous surgical scene segmentation methods. For the decoder, the following two key modules are proposed: low-rank matrix decomposition module (LMD) and boundary guided module (BGM) to solve the intraoperative complex noises and indistinguishable boundaries challenges in endoscopic images. The core idea of LMD is to separate the noise components from the useful feature information in the feature map through the low-rank matrix decomposition technique. In endoscopic surgical scenes, noise is generated in surgical images due to motion blur, blood splash, and water mist on the tissue surface caused by surgical operations. These noises reduce the segmentation accuracy of the network. LMD decomposes the feature map into a low-rank matrix containing the main feature information of the image and a sparse matrix containing the noise and outliers through nonnegative matrix factorization. Through this process, LMD can effectively remove the noise and provide a high-quality feature map for subsequent segmentation tasks. In surgical scenes, the boundaries between different tissues and instruments are highly indistinguishable due to contact, occlusion, or similar texture features. To solve this problem, BGM uses a combination of boundary-sensitive Laplace convolutions and normal convolutions to compute the boundary maps of the ground truth and the highest resolution feature maps, respectively. In addition, BGM uses a combination of cross-entropy loss function and dice loss function to guide the network to learn the boundary features and pay attention to the boundary region during training, thus improving the ability to recognize the boundary. A split laparoscopic surgical robot platform was constructed with the practical operational needs of the surgical process, integrating the advanced Lap Game endoscopic simulator, the Franka robotic arms, and a high-precision endoscopic imaging system, to apply the proposed MDVNet to actual surgical scenarios and verify its effectiveness. Users can manipulate the robotic arm equipped with surgical instruments through the control handle and complete endoscopic surgical operations such as cutting, freeing, and suturing in the endoscopic simulator to simulate the process of surgery. The segmentation results of the network are displayed on the user console to assist the surgeon in performing endoscopic surgery. **Result** To fully validate the effectiveness and potential of the proposed MDVNet, we subjected it to comprehensive comparative analysis with other advanced surgical scene segmentation methods on three different surgical scene datasets, namely, the robotic surgical scene dataset (Endovis2018), cataract surgical scene dataset (CaDIS), and minimally invasive laparoscopic surgery dataset (MILS). Experimental results show that MDVNet achieves the best segmentation results on all three datasets, with the intersection over union (mIoU) of 51.19% on Endovis2018, 71.28% on CaDIS (Task III), and 52.47% on MILS. Visualization results on the three datasets also illustrate that MDVNet can effectively segment multiple targets such as surgical instruments and anatomical areas in surgical scenes. We also conducted ablation experiments on the Endovis2018 dataset with three different modules, DLKA, LMD, and BGM. The results demonstrate that the different modules in MDVNet complement each other and can be combined to produce a positive gain effect on the whole method. Finally, the proposed MDVNet was employed on the laparoscopic surgical robot, and the segmentation results of the network were superimposed with the original surgical images for output to assist the surgeon in performing laparoscopic surgery. **Conclusion** To solve the three major challenges of endoscopic surgical scene segmentation, including target size variation, intraoperative complex noises, and indistinguishable boundaries, this work proposes an innovative surgical scene segmentation network named MDVNet. It is composed of three modules, DLKA, LMD, and BGM. For the encoder, DLKA can extract the multiscale features of different segmentation targets by applying multiscale large kernel attention and perform adaptive feature fusion by dynamic selection mechanism, effectively reducing the misidentification caused by the change of target sizes. For the decoder, LMD first fil-

ters out the noise in the feature maps and obtains high-quality feature maps. BGM guides the model to learn the boundary features of the surgical scene by calculating the loss between the boundary maps of the feature maps and the ground truth. MDVNet achieves state-of-the-art (SOTA) results on three different surgical scene datasets Endovis2018, CaDIS, and MILS. Code is available at https://github.com/YubinHan73/MDVNet.

Key words: endoscopic surgical robot; semantic segmentation; large kernel convolution; low-rank matrix decomposition (LMD); boundary segmentation

0 引　言

腔镜手术机器人是一种现代化的智能医疗设备。如图1所示，腔镜手术机器人一般由主端医生操作台、三维腹腔镜及摄像系统和从端病人操作台构成。腔镜手术机器人的摄像系统可以将手术视野真实地反映在医生操作台，医生通过在操作台上操作从端机械臂完成腔镜微创手术。相较传统手术方式，腔镜手术具有手术创伤小、恢复周期短和手术成功率高等特点，现已成为计算机辅助微创手术(robot-assisted minimally invasive surgery, MIS)的重要临床应用(Su等，2021；D'Souza等，2019)。

图1　腔镜手术机器人系统

Fig. 1　Endoscopic surgical robot system

理想的智能腔镜手术机器人能从视觉、听觉和力觉上为医生进行手术操作提供多模态感知信息(谢于廉，2022)，有效提高医生手术操作精准度，减少手术伤口及术后并发症。然而，现有腔镜手术机器人的多模态感知系统仍不完备，缺少高精度的手术场景解析和力反馈机制，手术效果高度依赖于医生的临床经验和技巧。若能开发有效的腔镜手术机器人场景分割系统，通过计算机视觉中的语义分割方法，把手术场景中的手术区域(解剖结构)和手术器械等对象的分割结果融合到医生腔镜视频视野中，显示手术靶标信息，将大幅增强医生识别定位手术目标的能力，进一步提高手术操作精度(王田苗等，2019)。此外，这项技术还能辅助执行许多术前和术后任务，包括手术风险评估(Collins等，2021)、手术导航(Lei等，2020)和手术技能评估(Liu等，2021)等。这些应用可以减轻医生的工作量，降低临床手术风险，对临床工作的开展具有重要意义。

精确解析腔镜手术图像的整个场景是一项极具挑战性的任务。与自然图像相比，医学影像具有分辨率低、对比度低和目标分散等特性(Despotović等，2015)。并且，腔镜手术在密闭的人体腔道完成，手术视野有限且视角不稳定。随着手术的进行，由于切割、牵引和吻合等手术操作，解剖结构如肾脏、小肠等的特征会发生显著变化。此外，手术器械和解剖结构的相互遮挡进一步增加了手术场景的复杂性(Liu等，2024)。

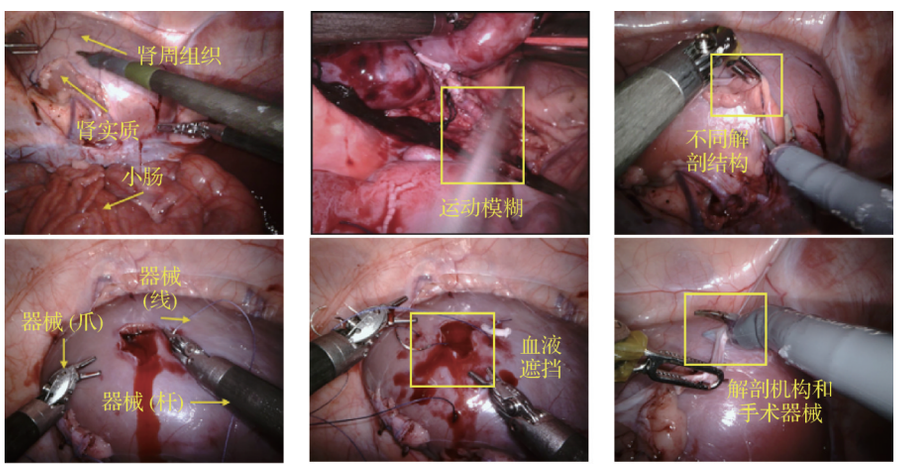
总体而言，腔镜手术场景分割主要面临术中目标尺寸差异大、术中多源噪声干扰和边界难以分辨3个挑战。如图2所示，结合机器人手术场景数据集Endovis2018(Allan等，2020)中的部分测试图像对手术场景分割的挑战进行详细说明。

1)目标尺寸差异大。腔镜手术场景分割的目标包括多种手术区域和手术器械。如图2(a)所示，肾实质、肾周组织(肾筋膜和肾周脂肪)和小肠的尺寸有较大差异，且特征会随着手术进行不断变化。同时，手术操作也涉及到多种不同尺寸的器械(杨磊等，2023)。分割目标的尺寸多样性对深度学习网络模型的分割性能和鲁棒性有很大的挑战(张学峰等，2023)。

2)术中多源噪声。在进行腔镜手术时，手术视野中会出现多种类型的噪声，这些噪声会对手术的成功率和安全性产生影响。由于腹腔镜图像采集频率和刷新率有限，在快速运动和操作中容易产生运动模糊。同时，手术过程中切除、切开和分离等操作会产生一定的出血情况，血液污染镜头表面会导致视野中出现遮挡。组织和器械之间的水分汽化和内

脏表面水汽凝结也会产生水雾遮挡。如图2(b)所示，在进行腔镜手术时，运动模糊和血液遮挡会改变分割目标的特征，影响医生视野。

3)边界难以分辨。由于边界特征多变且边界像素仅占手术视野的一小部分(花勇 等，2023)，精确分割不同区域的边界非常困难。如图2(c)所示，不同的解剖结构或手术器械相交或相互遮挡会产生曲折模糊的边界，现有的网络模型无法准确识别。



(a) 目标尺寸差异　(b) 术中多源噪声　(c) 边界难以分辨

图2　腔镜手术场景分割研究难点

Fig. 2　Main difficulties in surgical scene segmentation

((a) target size variation; (b) intraoperative complex noises; (c) indistinguishable boundaries)

在手术全场景分割方面，DeepLabV3+(Chen 等，2018)得到广泛应用。该网络通过带有空洞卷积的空间金字塔池化模块(atrous spatial pyramid pooling，ASPP)捕捉多尺度信息，在机器人手术场景分割挑战赛中取得了有竞争力的结果。HRNet(high-resolution network)(Wang 等，2021b)通过全程保持特征图的高分辨率表示来获取丰富的特征信息，在CaDIS(cataract dataset for image segmentation)(Gramatikopoulou 等，2021)白内障数据集中得到了最好的结果。Ni 等人(2022)提出了一种用于手术图像分割的双线性挤压推理网络，该方法采用高度池化和宽度池化在两个方向上挤压全局上下文，有助于解决解剖结构的局部相似性问题。但该工作没有关于手术过程中多源噪声对分割目标的干扰，这可能会导致分割目标的误识别或漏识别。Du 等人(2024)提出了一种多尺度特征金字塔聚合网络，通过密集乘法连接和局部注意力金字塔两个模块来聚合手术场景图像的多尺度注意力特征，有助于解决病理相似的问题。Jin 等人(2022)提出了一种基于时空 Transformer 和多源对比学习的新型框架 STswinCL(space-time shifted window contrastive learning)，该框架可对手术场景内和手术场景间的物体进行建模，利用不同帧之间的视频关系进行精确分割。然而，该网络模型有超过 100 M 参数，并且需要将视频片段作为输入，计算成本较高。徐旺旺等人(2024)提出了一种结合 Swin Transformer 和 UNet 的乳腺癌区域分割网络。Liu 等人(2024)结合腔镜手术场景的先验知识，提出了一种长条形大核卷积注意力网络(long strip kernel attention network，LSKA-Net)，通过不同拓扑结构的大核卷积，充分利用腔镜手术场景中解剖区域和手术器械的特征，在多个手术场景数据集上取得了先进的结果。但该模型参数量较大，运行速度有待提高。

为了实现精准的手术场景分割，本文提出了多尺度动态视觉网络(multi-scale dynamic visual network，MDVNet)。该网络采用编码器—解码器结构。编码器部分，在 SegNeXt(Guo 等，2022)中的多尺度大核卷积注意力的基础上，增加了动态选择机制进行特征的自适应融合，设计了动态大核卷积注意力模块(dynamic large kernel attention module，DLKA)，通过不同尺寸的大核卷积构建视觉注意力来提取不同分割目标的多尺度特征，提高分割准确率，有效解决分割目标尺寸差异大的问题。在解码器部分，该网络重点解决手术场景中的多源噪声和

边界难以分割问题。由于低秩矩阵分解可以重构特征图中隐藏的低秩矩阵，进而去除特征图中的噪声（Lu等，2014），因此，本文设计了低秩矩阵分解模块（low-rank matrix decomposition module，LMD）滤除特征图中的噪声。最后，参考LSKANet（Liu等，2024）中提出的边界损失函数，通过边界引导模块（boundary guided module，BGM）引导网络学习手术场景的边界特征。

1 手术场景分割网络

本节首先介绍MDVNet的整体架构，然后介绍动态大核卷积注意力模块DLKA、低秩矩阵分解模块LMD和边界引导模块BGM等3个主要模块。为方便说明，首先定义以下表达。给定一个腔镜手术场景图像$\boldsymbol{I} \in \mathbf{R}^{C \times H \times W}$，其中$C$, H, W分别为通道尺寸、图像高度和宽度，GT代表手术图像$\boldsymbol{I}$对应的标签。

1.1 网络概述

本文所提出的MDVNet的网络架构如图3所示，网络采用编码器—解码器结构。

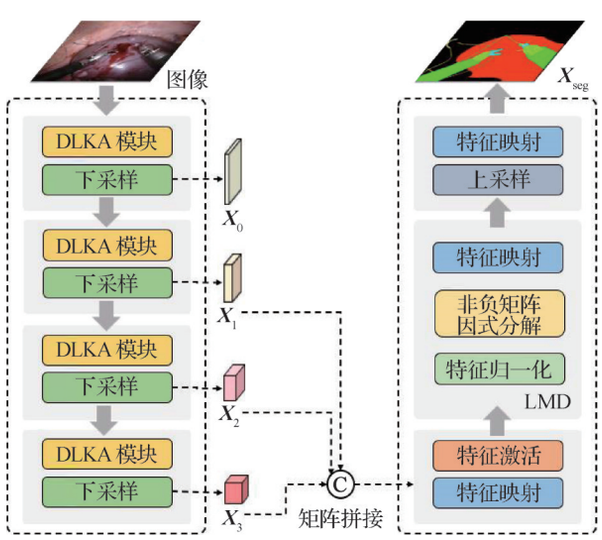


图3 MDVNet网络架构

Fig. 3 The framework of the proposed MDVNet

给定一个腔镜手术场景图像$\boldsymbol{I} \in \mathbf{R}^{C \times H \times W}$，$\boldsymbol{I}$首先通过编码器进行特征编码获得特征图。编码器由多次重复的动态大核卷积注意力模块DLKA和前馈网络组成。其中，DLKA通过不同尺度的大核卷积分支对多个分割目标进行联合建模，并通过动态选择机制让网络自适应地提取手术场景特征。图像经过特征编码获得4个不同尺寸的特征图$\boldsymbol{X}_i \in \mathbf{R}^{C_i \times H_i \times W_i}$，其中$i \in \{0, 1, 2, 3\}$代表第$i$个阶段。解码器由特征映射和低秩矩阵分解模块LMD构成。其中，LMD通过低秩矩阵分解滤除手术场景中运动模糊和水雾、血液遮挡等噪声，最后通过特征映射获得分割结果$\boldsymbol{X}_{\text{seg}}$。在特征解码阶段，通过添加边界引导模块BGM计算出分割图$\boldsymbol{X}_0$的边界图$\boldsymbol{BM}$，并通过与标签的边界图$\boldsymbol{BM}_{\text{gt}}$计算损失函数$L_{\text{bd}}$。网络的整体损失函数由分割损失函数$L_{\text{seg}}$和边界损失函数$L_{\text{bd}}$组合构成。整个网络通过梯度下降法进行端到端训练。BGM只用在网络的训练阶段，在进行网络推理时不使用。

1.2 编码器

腔镜手术场景分割的目标包括多种手术区域和手术器械，分割目标尺寸范围广泛，从细小的手术针到较大的组织结构，这种尺寸上的差异对分割算法的鲁棒性有较大挑战（Shen等，2023），算法必须能够同时处理极小和极大的目标，而不丢失细节或产生误判。因此，高精度的手术场景分割算法需要有足够大的有效感受野，并可以综合手术场景特征进行联合建模。

1.2.1 动态大核卷积注意力

为解决不同分割目标尺寸差异大的问题，本文设计了动态大核卷积注意力DLKA进行特征建模，其结构如图4所示。输入的特征图$\boldsymbol{X}_i \in \mathbf{R}^{C_i \times H_i \times W_i}$首先经过批归一化和一个空间卷积层来聚合结构信息，该过程的数学表达为

$$\boldsymbol{X}_i' = Conv_{5\times 5}^{dw}\left(BN\left(\boldsymbol{X}_i\right)\right) \tag{1}$$

式中，$BN(\cdot)$代表批归一化操作，$Conv_{m\times n}^{dw}$代表卷积核大小为$m \times n$的深度可分离卷积。

DLKA采用3个不同尺寸的大核卷积注意力分支来计算特征图，从而使有效感受野的规模逐步扩大，让网络能够捕捉到更精细、信息量更大的特征。DLKA使用一对带有条形卷积核的深度可分离卷积层取代标准的大核卷积，可以减少使用大核卷积带来的计算开销。对于每个大核卷积分支j，分支注意力图$\boldsymbol{A}_{i,j}'$计算为

$$\boldsymbol{A}_{i,j}' = Conv_{k_j\times 1}^{dw}\left(Conv_{1\times k_j}^{dw}\left(\boldsymbol{X}_i'\right)\right) \tag{2}$$

为了充分利用不同尺度的视觉和结构信息，为每个注意力分支设定不同大小的卷积核$k_j = \{7, 11, 21\}$，其中$j \in \{0, 1, 2\}$。接下来，沿特征图的通道维并行使用平均值池化（average pooling，AP）和

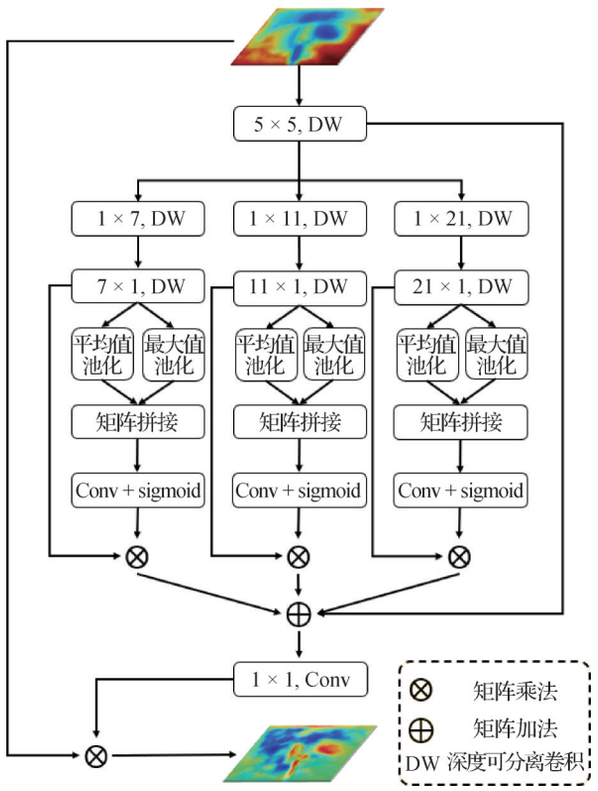


图 4　DLKA 结构

Fig. 4　The framework of the proposed DLKA

最大值池化(maximum pooling, MP)对特征图进行联合建模。之后,将两个特征权重沿通道维度进行拼接,该过程的数学表达为

$$\boldsymbol{w}_{i,j}^{\text{ave}} = AP\left(\boldsymbol{A}_{i,j}'\right) \tag{3}$$

$$\boldsymbol{w}_{i,j}^{\max} = MP\left(\boldsymbol{A}_{i,j}'\right) \tag{4}$$

$$\boldsymbol{w}_{i,j}' = f_{\text{Concat}}\left(\boldsymbol{w}_{i,j}^{\text{ave}}, \boldsymbol{w}_{i,j}^{\max}\right) \tag{5}$$

式中,$\boldsymbol{w}_{i,j}^{\text{ave}}$、$\boldsymbol{w}_{i,j}^{\max} \in \mathbf{R}^{1\times H_i\times W_i}$,$\boldsymbol{w}_{i,j}' \in \mathbf{R}^{2\times H_i\times W_i}$,$AP(\cdot)$ 和 $MP(\cdot)$分别代表平均值池化和最大值池化,Concat指的是矩阵沿着通道维度进行拼接操作。接下来,通过卷积和sigmoid激活函数进行特征映射和激活,得到分支的特征权重 $\boldsymbol{w}_{i,j}$,具体为

$$w_{i,j} = Conv_{1\times 1}\left(f_{\text{sigmoid}}\left(\boldsymbol{w}_{i,j}'\right)\right) \tag{6}$$

式中,$Conv_{1\times 1}$代表 1 × 1 卷积层。

分支注意力图由分支特征权重和阶段注意力图相乘得到,不同的注意力分支可以自适应地学习不同尺度的手术场景特征

$$\boldsymbol{A}_{i,j} = \boldsymbol{A}_{i,j}' \otimes \boldsymbol{w}_{i,j}' \tag{7}$$

式中,⊗代表矩阵乘法。3 个分支注意力图相加后与原始输入相加,最后通过 1 × 1 卷积层进行通道压缩得到对应的注意力图,该过程的数学表达为

$$\boldsymbol{Att}_i = Conv_{1\times 1}\left(\sum_{j=0}^{2}\boldsymbol{A}_{i,j} + \boldsymbol{X}_i'\right) \tag{8}$$

在得到注意力图 $\boldsymbol{Att}_i$ 后,将其与输入 $\boldsymbol{X}_i$ 相乘后通过前馈网络(feed forward network, FFN)得到特征增强之后的特征图,前馈网络FFN由卷积层、线性激活层和dropout层构成,计算过程为

$$\hat{\boldsymbol{X}}_i = FFN(\boldsymbol{Att}_i \otimes \boldsymbol{X}_i) \tag{9}$$

式中,$L_i = \{3, 3, 12, 3\}$。在编码器的 4 个阶段,连续使用 L_i 个 DLKA 进行特征编码。

1.2.2　卷积下采样

在得到 $\hat{\boldsymbol{X}}_i$ 后,通过卷积下采样进行特征升维。卷积下采样模块包含一个步长为 2、卷积核大小为 3 × 3 的卷积层和一个批归一化层。具体为

$$\boldsymbol{X}_{i+1} = BN(Conv_{3\times 3}(\hat{\boldsymbol{X}}_i)) \tag{10}$$

最终,编码器生成 4 个不同尺寸的特征图,分别为输入图像的 1/4、1/8、1/16 和 1/32。

1.3　解码器

不同于其他语义分割网络的解码器,本文提出的解码器只对后 3 个阶段的特征图进行特征映射和拼接。这是由于低维的特征图包含许多冗余的信息,会影响网络的分割性能。该过程的数学表达式为

$$\boldsymbol{X}_{\text{out}}' = f_{\text{Concat}}\left(\boldsymbol{X}_1, UP_{\times 2}(\boldsymbol{X}_2), UP_{\times 4}(\boldsymbol{X}_3)\right) \tag{11}$$

$$\boldsymbol{X}_{\text{out}} = Conv_{1\times 1}\left(\boldsymbol{X}_{\text{out}}'\right) \tag{12}$$

式中,$UP_{\times n}$ 表示线性上采样 n 倍。

1.3.1　低秩矩阵分解

矩阵中最大的不相关的向量的个数,称为秩(rank)。在图像处理中,秩可以理解为图像所包含的信息的丰富程度,如果图像的秩比较高,往往是因为图像中的噪声比较严重。以网络输出的特征图 $\boldsymbol{X}' \in \mathbf{R}^{C'\times H'\times W'}$ 为例,将张量 $\boldsymbol{X}'$ 展开为 $\overline{\boldsymbol{X}'} \in \mathbf{R}^{C'\times H'W'}$。在图像中,由于像素是线性相关的,说明 $\overline{\boldsymbol{X}'}$ 中的每个像素都可以表示为元素小于 $H'W'$ 的基准的线性组合。因此,通过低秩矩阵分解的方式重构特征图中的隐藏的低秩矩阵,进而去除特征图中的噪声(Lu等,2014)。

假设存在一个矩阵 $\boldsymbol{M} \in \mathbf{R}^{d\times n}$,则会存在矩阵 $\boldsymbol{D} \in \mathbf{R}^{d\times r}$ 和 $\boldsymbol{C} \in \mathbf{R}^{r\times n}$,满足以下表达式

$$\boldsymbol{M} = \bar{\boldsymbol{M}} + \boldsymbol{N} = \boldsymbol{D}\times\boldsymbol{C} + \boldsymbol{N} \tag{13}$$

式中,$\bar{\boldsymbol{M}} \in \mathbf{R}^{d\times n}$ 是输出的低秩重构矩阵,$\boldsymbol{N} \in \mathbf{R}^{d\times n}$ 是要丢弃的噪声矩阵。这里 $\bar{\boldsymbol{M}}$ 具有低秩属性,满足

$$rank(\bar{\boldsymbol{M}}) \leqslant \min(rank(\boldsymbol{D}), rank(\boldsymbol{C})) \leqslant r \quad (14)$$

利用式(13)的矩阵分解模型重构矩阵$\boldsymbol{M}$中的低秩部分$\bar{\boldsymbol{M}}$,并丢弃噪声部分$\boldsymbol{N}$,可以过滤掉原始矩阵中冗余和不完整的信息。低秩矩阵分解模块LMD的结构如图5所示。整个过程的数学表达式为

$$\boldsymbol{M} = Conv_{1\times1}\left(ReLU(\boldsymbol{X}_{\text{out}})\right) \quad (15)$$

式中,$ReLU(\cdot)$的作用是保证输入矩阵非负,之后使用非负矩阵因式分解(non-negative matrix factorization,NMF)重建低秩矩阵$\bar{\boldsymbol{M}}$,并丢弃噪声矩阵$\boldsymbol{N}$。迭代规则选择乘法更新规则,迭代次数设置为10。

之后,通过特征映射和跳跃连接得到LMD的输出$\bar{\boldsymbol{X}}_{\text{out}}$,表达式为

$$\bar{\boldsymbol{X}}_{\text{out}} = ReLU\left(\boldsymbol{X}_{\text{out}} + \left(Conv_{1\times1}(\bar{\boldsymbol{M}})\right)\right) \quad (16)$$

通过低秩分解模块,对手术场景特征进行去噪处理,获得了高质量的特征图。

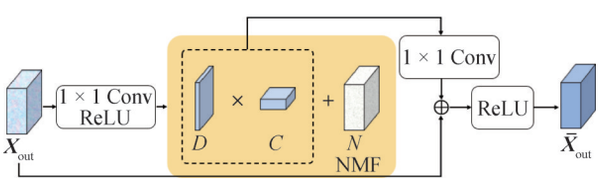


图5 LMD结构

Fig. 5 The framework of the proposed LMD

1.3.2 边界引导

腔镜手术涉及各种解剖组织和器械之间的交互。随着手术的进行,由于切割、牵引或其他操作,解剖区域的特征会发生显著变化,使得边界更加难以区分。因此,本文设计了边界引导分割模块BGM,帮助网络分割难以分辨的目标边界。

如图6所示,使用3×3的拉普拉斯卷积(Laplace convolution,Lap.Conv)来计算$\boldsymbol{GT}$的边界图,并使用3个不同分辨率的分支来获得丰富的边界信息,具体为

$$\boldsymbol{BM}_{\text{gt}}^{i} = UP_{\times S_i}\left(Lap_{3\times3}^{S_i}(\boldsymbol{GT})\right) \quad (17)$$

式中,$Lap_{3\times3}^{S_i}$代表3×3的拉普拉斯卷积,步长设置为S_j。对于$i \in \{0, 1, 2\}$,S_j设置为1,2,4。

之后,3个分支的边界特征图进行加权求和得到$\boldsymbol{GT}$的边界图,具体为

$$\boldsymbol{BM}_{\text{gt}} = \sum_{i=0}^{2} \alpha_i \times \boldsymbol{BD}_{\text{gt}}^{i} \quad (18)$$

式中,α_i是用于组合边界特征图的权重,α_i设置为0.6,0.3和0.1。$\boldsymbol{BM}_{\text{gt}} \in \mathbf{R}^{1\times H\times W}$为$\boldsymbol{GT}$的边界图。

接下来,计算最高分辨率特征图$\boldsymbol{X}_0$的边界图$\boldsymbol{BM}$,具体为

$$\boldsymbol{BM} = UP_{\times4}(\Gamma(\boldsymbol{X}_0^{c})) \quad (19)$$

式中,$\Gamma(\cdot)$是1×1卷积层、线性激活层、批归一化和1×1卷积层的组合。

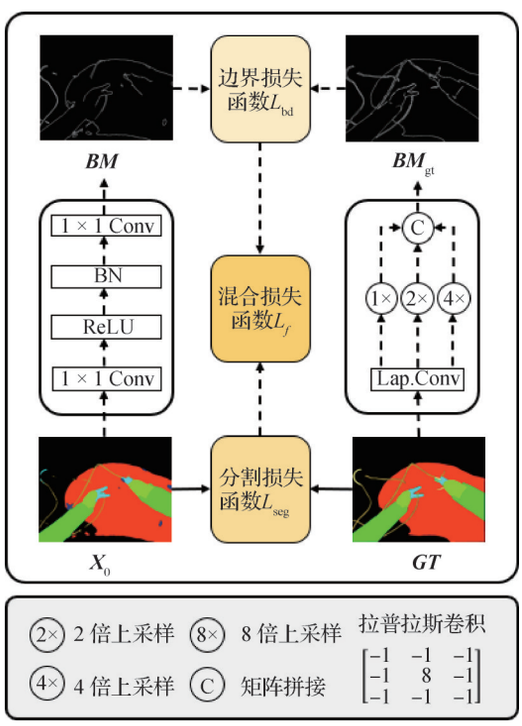


图6 BGM结构

Fig. 6 The framework of the proposed BGM

1.3.3 损失函数

LMD输出的特征图$\bar{\boldsymbol{X}}_{\text{out}}$首先经过dropout层和卷积层进行通道映射,之后通过上采样得到分割图$\boldsymbol{X}_{\text{seg}}$。整个过程的数学表达式为

$$\boldsymbol{X}_{\text{seg}} = UP_{\times4}(Conv_{1\times1}^{\text{seg}}(\bar{\boldsymbol{X}}_{\text{out}})) \quad (20)$$

本文采用交叉熵损失函数(cross-entropy loss)L_{ce}计算分割结果与标签之间的分割损失L_{seg},具体为

$$L_{\text{seg}} = L_{\text{ce}}(\boldsymbol{X}_{\text{seg}}, \boldsymbol{GT}) \quad (21)$$

边界分割是典型的类别不平衡问题。因此,使用交叉熵损失函数L_{ce}和Dice损失函数L_{dice}的组合来设计边界损失函数L_{bd},表达式为

$$L_{\text{bd}} = L_{\text{ce}}\left(\boldsymbol{BM}, \boldsymbol{BM}_{\text{gt}}\right) + L_{\text{dice}}(\boldsymbol{BM}, \boldsymbol{BM}_{\text{gt}}) \quad (22)$$

为引导网络精准解析手术场景,本文使用L_{seg}和L_{bd}的线性组合作为MDVNet的整体损失函数L_f。具体为

$$L_f = \alpha L_{\text{seg}} + (1-\alpha)L_{\text{bd}} \quad (23)$$

式中,α为权重系数,本文将其设置为0.5。值得注

意的是，BGM 只在训练过程中应用，在推理阶段被舍弃以提高推理速度。该模块可以在几乎不增加额外计算成本的情况下有效改善网络模型的分割结果。

2 实验结果与分析

本节将提出的 MDVNet 与其他先进的手术场景分割方法在 3 个不同的手术场景数据集 Endovis2018，CaDIS（cataract dataset for image segmentation）和 MILS（minimally invasive laparoscopic surgery dataset）上进行了实验，并对实验结果进行分析。此外，本节对 MDVNet 进行了消融实验，以验证不同模块对网络性能的影响。

2.1 实验数据

1）机器人手术场景数据集 Endovis2018。Endovis2018 在 2018 年 MICCAI（Medical Image Computing and Computer Assisted Intervention Society）机器人场景分割挑战赛中发布。该数据集由 19 个腔镜手术视频组成，划分为 15 个训练视频，共 2 235 帧手术场景图像以及 4 个测试视频，共 997 帧手术场景图像，数据来自使用专用硬件在 da Vinci X 或 Xi 系统上记录的单个猪手术过程，图像分辨率为 1 280 × 1 024 像素。该数据集任务是将腔镜手术场景分为 12 个类别，包括背景、3 类解剖区域和 8 类医疗设备。

2）白内障手术数据集 CaDIS。CaDIS 由 25 个白内障手术视频组成，共有 3 550、534 和 586 帧用于训练、验证和测试，每帧分辨率为 960 × 540 像素。数据集为解剖结构、手术器械和其他物体提供像素级注释，并定义了 3 个颗粒度逐渐增大的子任务。任务Ⅰ涉及 8 个分割目标，包括 4 个解剖结构、3 个混合类别（手术胶带、手和瞳孔牵引器）和 1 个包含所有仪器的类别。任务Ⅱ包括器械分类，并将分割目标数量增加到 17 个。手术器械根据外观相似性进行分类，最终形成 9 个不同的类别。任务Ⅲ对手术器械进行更细化的分类，将每种器械的尖端和手柄作为单独的类别，包含 25 个分割目标。

3）腹腔镜微创手术数据集 MILS。腹腔镜微创手术数据集（minimally invasive laparoscopic surgery dataset，MILS）是一个临床私人数据集。数据采集自湖南省某三甲医院提供的 8 名患者的全直肠系膜切除（total mesorectal excision，TME）手术的视频。TME 手术是直肠癌的标准手术方式，其要求在正确的层面完整切除直肠及其系膜，以避免损伤血管和直肠间的神经等脆弱组织。该数据集包含 3 000 帧不同阶段的腔镜手术场景图像，其中 2 250 帧用于训练，750 帧用于测试。每帧分辨率为 2 048 × 1 024 像素，包括 8 个类别（1 个背景类别、1 个组织间隙和 6 种不同的手术器械）。该数据集中需要分割的解剖结构是组织间隙，这是一种典型的非刚性组织，这使得 MILS 与其他的手术场景数据集有所不同。

2.2 实验环境与设置

1）实验设置。本节中所有实验均基于 PyTorch 深度学习框架和单个 NVIDIA RTX A6000 GPU 进行训练和推理。Endovis2018 和 MILS 中的图像调整为 768 × 768 像素，CaDIS 中的图像尺寸调整为 768 × 512 像素。使用 MMSegmentation 作为代码框架，并使用其提供的预训练权重来训练所有模型。所有训练的模型都采用了相同的数据增强方法，包括随机调整大小、随机裁剪、随机翻转和光度失真以减少过拟合。网络采用 AdamW 优化器更新网络参数以及最小化损失函数，并设置初始学习率为 6×10^{-5}。网络采用 warm-up 策略训练 200 个 Epoch，批次大小（batchsize）设置为 8。

2）评价指标。为了综合评估所提出的方法，对于 Endovis2018、CaDIS 和 MILS，本文使用语义分割的通用指标进行评价，分别为像素准确率（pixel accuracy，PA）、平均交并比（mean intersection over union，mIoU）和平均像素准确率（mean pixel accuracy，mPA）。

2.3 对比实验分析

首先，本节将提出的 MDVNet 与各种优秀的语义分割方法在 Endovis2018、CaDIS 和 MILS 3 个数据集上进行比较。比较的方法分为 3 组：第 1 组包括 4 种多尺度特征融合方法 U-Net（Ronneberger 等，2015）、DeepLabv3+（Chen 等，2018）、UPerNet（unified perceptual parsing network）（Xiao 等，2018）和 HRNetV2（high-resolution network v2）（Wang 等，2021b），它们是 Endovis2018 或 CaDIS 公开发表论文中使用的网络模型；第 2 组包括另外 3 种经典的多尺度特征融合方法 OCRNet（object-contextual representations network）（Yuan 等，2020）、STDCNet（short term dense concatenate network）（Fan 等，2021）和 PSPNet（pyramid scene parsing network）（Zhao 等，

2017)；第3组包括3种最新的语义分割方法：基于分层Transformer结构的语义分割网络SegFormer(Xie等，2021)、基于大核卷积注意力的语义分割网络SegNeXt(Guo等，2022)和在腔镜手术图像分割领域公开发表的最优方法LSKANet(long strip kernel attention network)(Liu等，2024)。接下来对实验结果进行展示和分析。

1)Endovis2018。表1展示了不同分割方法在Endovis2018数据集上的结果。在对比的3组方法中，LSKANet通过两种不同拓扑结构(串联和并联)的大核卷积注意力分别提取解剖结构和手术器械的特征，充分利用手术场景的视觉和结构信息，取得了较优的分割结果，mIoU达到了50.92%。值得注意的是，本文所提出的MDVNet达到了51.19% mIoU、85.12% PA和61.58% mPA，在该数据集上取得了最好的结果。

为了更好地说明所提方法的效果，本文将部分手术图像的分割结果进行了可视化，如图7所示。从可视化结果可以看出，MDVNet可以有效识别不同尺寸的目标，如第1行中的大目标(肾实质)和小目标(手术针)。相较于其他的网络模型，MDVNet对于解剖结构的边界识别更加准确，如第2行肾实质和肾周组织的边界。MDVNet还能对存在噪声干扰的图像进行准确分割，如第3行中出现运动模糊的器械和发生畸变的视野边缘。

表1　不同方法在Endovis2018上的对比实验结果

Table 1　Comparison experimental results of different methods on Endovis2018

/%

方法	mIoU	PA	mPA
UNet(Ronneberger等，2015)	37.22	74.75	48.71
DeepLabv3+(Chen等，2018)	44.50	81.36	54.66
UPerNet(Xiao等，2018)	45.85	80.16	55.58
HRNetV2(Wang等，2021b)	46.64	81.04	59.76
OCRNet(Yuan等，2020)	48.09	82.23	59.67
STDCNet(Fan等，2021)	45.24	82.24	56.56
PSPNet(Zhao等，2017)	45.81	80.63	55.21
SegFormer(Xie等，2021)	49.36	82.62	60.16
SegNeXt(Guo等，2022)	50.34	83.66	59.82
LSKANet(Liu等，2024)	50.92	84.47	60.75
MDVNet(本文)	**51.19**	**85.12**	**61.58**

注：加粗字体表示各列最优结果。

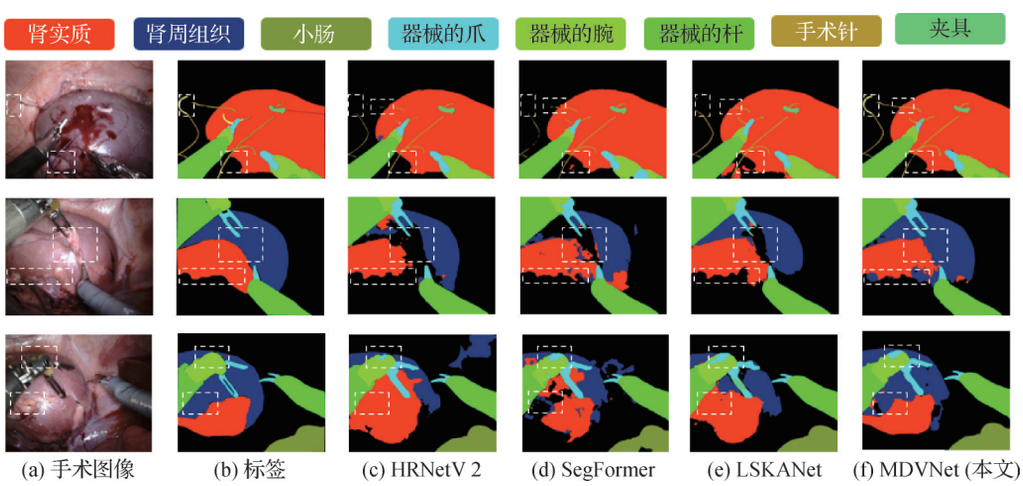


图7　不同方法在Endovis2018上的可视化结果

Fig. 7　Visualization results of different methods on Endovis2018

((a) surgical images; (b) labels; (c) HRNetV2; (d) SegFormer; (e) LSKANet; (f) MDVNet(ours))

2)CaDIS。本文在白内障手术数据集CaDIS上对比了MDVNet和其他方法，实验结果如表2所示。SegFormer的编码器采用了分层Transformer结构，将输入图像分成多个阶段处理，在每个阶段中，前一层的特征图都会通过合并重叠特征图的方式生成新的特征图，各层级特征之间的交互融合使得SegFormer拥有较大的有效感受野，在CaDIS的3个任务上都取得了非常有竞争力的结果。本文提出的MDVNet在该数据集上表现出色，在所有任务上都获得了最佳的mIoU指标，分别为86.03%、

表 2　在 CaDIS 上的对比实验结果

Table 2　Comparison experimental results on CaDIS

/%

方法	任务 Ⅰ (8类)			任务 Ⅱ (17类)			任务 Ⅲ (25类)		
	mIoU	PA	mPA	mIoU	PA	mPA	mIoU	PA	mPA
UNet(Ronneberger 等,2015)	83.32	94.43	88.89	70.21	94.34	80.23	59.33	94.14	71.64
DeepLabv3+(Chen 等,2018)	85.63	94.82	91.54	75.84	**94.74**	84.92	65.93	94.74	78.09
UPerNet(Xiao 等,2018)	85.94	94.73	91.68	77.04	94.73	84.93	70.11	94.43	80.45
HRNetV2(Wang 等,2021b)	85.52	94.54	91.59	76.25	94.53	85.10	68.00	94.50	78.60
OCRNet(Yuan 等,2020)	85.48	94.79	90.88	76.28	94.52	85.38	70.17	94.60	80.80
STDCNet(Fan 等,2021)	84.02	84.40	91.00	73.75	94.21	84.28	64.48	94.22	76.52
PSPNet(Zhao 等,2017)	85.07	94.64	84.27	75.58	94.68	84.16	66.05	94.43	77.48
SegFormer(Xie 等,2021)	85.38	94.48	91.38	77.74	94.40	**86.28**	70.22	94.35	81.09
SegNeXt(Guo 等,2022)	85.83	94.77	**91.92**	77.51	94.62	85.67	70.04	94.64	80.46
LSKANet(Liu 等,2024)	86.02	**94.97**	91.28	77.97	94.70	85.40	71.02	94.54	81.24
MDVNet(本文)	**86.03**	94.80	91.66	**78.18**	94.73	86.00	**71.28**	**94.65**	**81.46**

注:加粗字体表示各列最优结果。

78.18% 和 71.28%。并在总共 9 项评价指标(3 项任务各 3 项指标)中的 5 项指标上取得了最好的结果。与次优模型相比,MDVNet 在更精细任务上获得了更大的提高(mIoU 提升更为明显)。这些结果进一步证明了本文方法的有效性。从表 2 中可以看出,MDVNet 在任务 Ⅰ 和任务 Ⅱ 的 PA 和 mPA 并没有取得最优,这可能是因为前两个任务细粒度较低,存在类别不平衡问题。像素占比较大的分割目标的正确率对 PA 和 mPA 两项分割指标影响较大。

如图 8 所示,本文同样可视化了部分模型在 CaDIS(任务 Ⅲ)上的结果,黑色虚线框代表网络模型误识别的部分。从结果可以看出,MDVNet 可以很好地分割多个目标相交形成的边界,在特征扭曲的视场边缘也表现稳定,如第 2 行和第 3 行黑色虚线框的部分。同时,在第 1 行黑色虚线框的部分视野被医生手部遮挡的情况下,所提出的模型仍能准确分割出不同目标的边界。

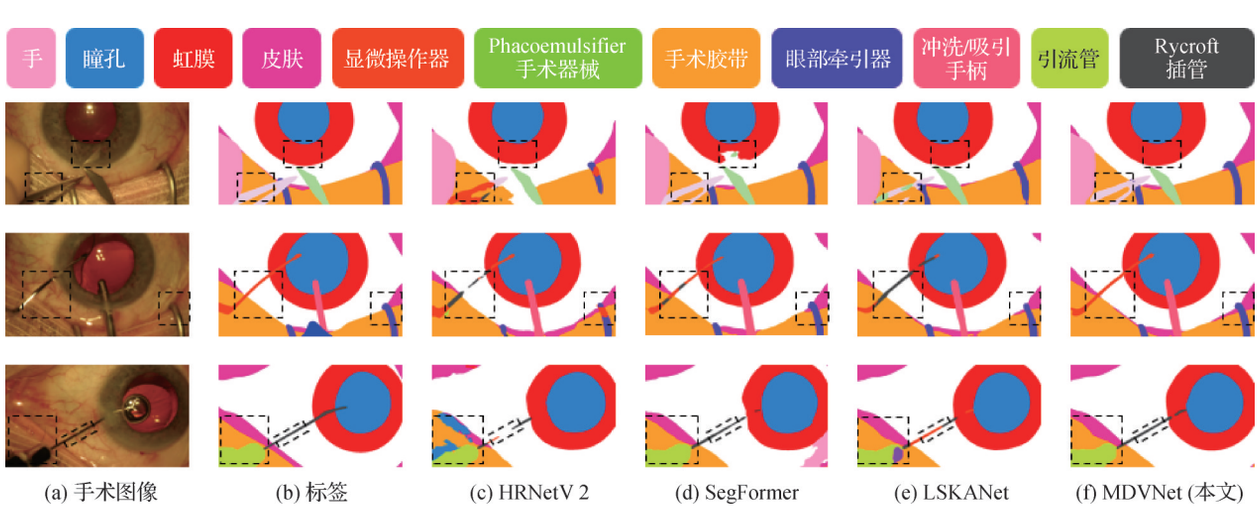


图 8　不同方法在 CaDIS 上的可视化结果

Fig. 8　Visualization results of different methods on CaDIS

((a) surgical images; (b) labels; (c) HRNetV2; (d) SegFormer; (e) LSKANet; (f) MDVNet(ours))

3)MILS。不同网络模型在MILS上的实验结果如表3所示。在对比的3组方法中,OCRNet通过建模像素与对象区域之间的关系,在该数据集中取得了次优结果。值得注意的是,所提出的MDVNet在MILS中取得了SOTA(state-of-the-art)结果,mIoU为52.47%,PA为95.59%,mPA为71.94%。这些结果表明本文方法即使在困难样本上也具有出色的分割性能。

部分网络模型在MILS上的可视化结果如图9所示。可视化结果表明,MDVNet能够准确识别具有相似特征的手术器械,如可以精准地分割第1行手术图像中的3种不同器械的边界,其他网络模型无法正确识别肠钳这一目标。此外,对于组织间隙这种非刚性的目标,本文方法也能精确定位并分割其边界,如图9中第2行和第3行中的组织间隙,MDVNet是唯一可以进行准确分割的模型。

2.4 消融实验分析

为了测试MDVNet网络中每个模块对网络性能的影响和网络设计的有效性,本文对DLKA、LMD和BGM 3个不同模块进行了全面的消融实验和结果分析。消融实验在Endovis2018数据集上进行。

1)DLKA的消融实验。为验证DLKA中不同模

表3 在MILS上的对比实验结果

Table 3 Comparison experimental results on MILS

/%

方法	mIoU	PA	mPA
UNet(Ronneberger等,2015)	35.62	93.44	56.69
DeepLabv3+(Chen等,2018)	46.67	95.12	67.65
UPerNet(Xiao等,2018)	45.70	94.76	63.22
HRNetV2(Wang等,2021b)	48.14	95.13	68.14
OCRNet(Yuan等,2020)	48.70	94.87	67.71
STDCNet(Fan等,2021)	47.38	95.16	68.35
PSPNet(Zhao等,2017)	45.84	95.09	64.42
SegFormer(Xie等,2021)	46.70	95.11	66.98
SegNeXt(Guo等,2022)	47.86	95.35	69.90
LSKANet(Liu等,2024)	52.09	95.58	70.45
MDVNet(本文)	**52.47**	**95.59**	**71.94**

注:加粗字体表示各列最优结果。

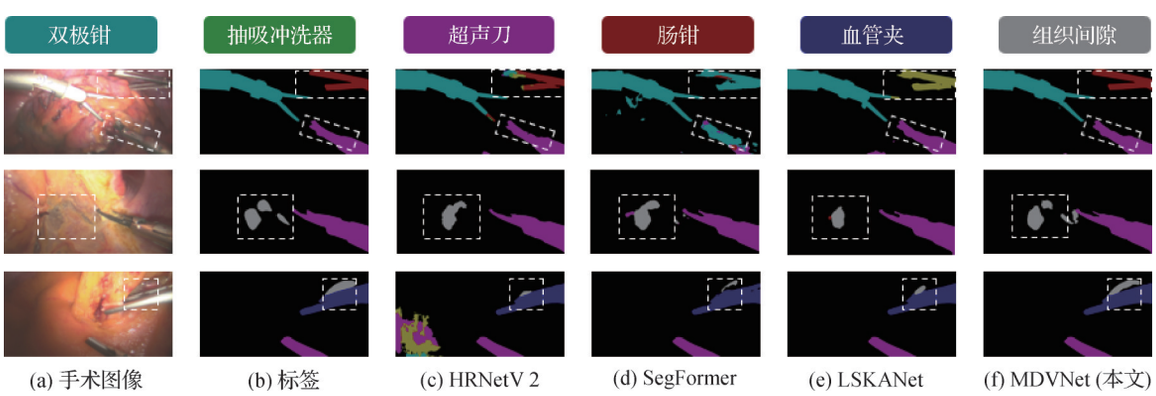


图9 不同方法在MILS上的可视化结果

Fig. 9 Visualization results of different methods on MILS

((a) surgical images; (b) labels; (c) HRNetV2; (d) SegFormer; (e) LSKANet; (f) MDVNet(ours))

块的大核卷积设计和动态权重设计的有效性，本文构建了3个模型，使用普通的3×3卷积代替大核卷积构建了模型1；使用大核卷积分支而不使用动态权重设计构建了模型2；同时使用大核卷积设计和动态权重设计构建了模型3，实验结果如表4所示。可以看出，单独使用大核卷积注意力（模型2）也可以提高网络的性能，这是由于使用大核卷积构建注意力可以扩大模型有效感受野，进行长距离特征建模。在模型2的基础上添加动态权重模块（模型3）可以为不同的大核卷积分支构建权重，让网络自适应地学习不同尺度的手术场景特征。添加完整的DLKA相较于模型1在Endovis2018数据集上得到了3.09% mIoU和3.88% mPA的提升。

2）解码器的消融实验。在网络的解码器中，本文设计了低秩矩阵分解模块LMD和边界引导模块BGM。为了验证解码器中两个模块的有效性，本文构建了3个模型，分别是模型3（不添加LMD和BGM）、模型4（只添加LMD）和模型5（只添加BGM），实验结果如表5所示。从表中可以看出，LMD和BGM都能提升模型的性能。当添加所有模块时，完整的MDVNet取得了51.59% mIoU和61.58% mPA，这说明所提出的不同模块之间是互补的，并且组合起来能够对网络模型产生正向的增益效果。

3）LMD的噪声消融实验。在进行腔镜手术时，手术视野中会出现多种类型的噪声，主要包括运动模糊和水雾遮挡等，这些噪声会对手术的成功率和安全性产生影响。为了验证低秩矩阵分解模块LMD可以有效处理术中多源噪声的干扰，本文参考Hendrycks和Dietterich（2019）的工作，给机器人手术场景数据集Endovis2018添加了运动模糊和水雾两种噪声，样本示例如图10所示。

本节在添加噪声的数据上对低秩矩阵分解模块LMD进行消融实验，实验结果如表6所示。完整的MDVNet在运动模糊和水雾遮挡两种噪声的情况下，mIoU分别提升了1.62%和4.78%，这说明了LMD处理术中多源噪声的有效性。

为了更好地说明所提出方法的效果，本文将部分手术图像的分割结果进行了可视化，结果如图11所示。从可视化结果可以看到，运动模糊和水雾对网络模型的干扰很大，尤其是对边界区域的分割。本文提出的MDVNet在有噪声干扰的情况下仍能很好地分割出不同目标的轮廓，误识别的情况较少。而没有添加LMD的网络模型（模型5）或无法完整覆盖目标区域，或多余分割了不属于目标的区域，受噪

表4　DLKA消融实验结果

Table 4　Ablation experiment results of DLKA

模型	模块		Endovis2018	
	大核卷积	动态权重	mIoU/%	mPA/%
模型1	–	–	46.46	57.56
模型2	√	–	48.39	59.04
模型3	√	√	**49.55**	**61.44**

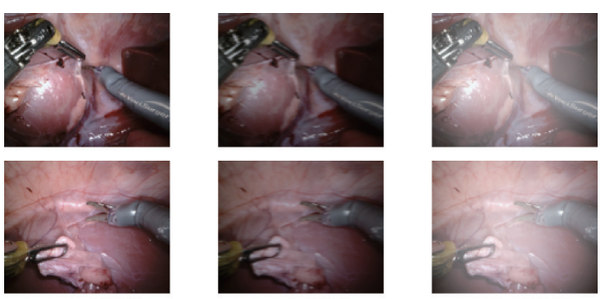
注：加粗字体表示各列最优结果。“√”表示保留相应的模块，“–”表示不保留。

表5　解码器中不同模块的消融实验结果

Table 5　Ablation experiment results of modules in encoder

模型	模块		Endovis2018	
	LMD	BGM	mIoU/%	mPA/%
模型3	–	–	49.55	60.83
模型4	√	–	50.96	61.44
模型5	–	√	50.57	61.28
MDVNet（本文）	√	√	**51.59**	**61.58**

注：加粗字体表示各列最优结果。“√”表示保留相应的模块，“–”表示不保留。



(a) 手术图像　(b) 运动模糊　(c) 水雾遮挡

图10　添加噪声的Endovis2018的样本示例

Fig. 10　Image examples of Endovis2018 with added noises ((a) surgical images; (b) motion blur; (c) water mist)

表6　LMD的噪声消融实验结果

Table 6　Ablation experiment results of LMD in processing noises

模型	模块	Endovis2018（mIoU）	
	LMD	运动模糊/%	水雾遮挡/%
模型5	–	35.92	42.76
MDVNet（本文）	√	**37.54**	**47.54**

注：加粗字体表示各列最优结果。“√”表示保留相应的模块，“–”表示不保留。

声影响较大。这进一步说明了LMD处理术中多源噪声的有效性。

2.5 模型参数分析

表7显示了不同分割模型和本文提出的MDVNet的参数量(Params)、浮点运算量(floating point operations per second, FLOPs)和网络推理速度(frame per second, FPS)。图像尺寸设置为768×512像素,批次大小(batchsize)设置为1。

从表1、表2、表3和表7可以看出,MDVNet在3个不同场景的手术数据集上取得了最好的结果,且相较于SOTA模型LSKANet,参数量减少了62.83%,浮点运算量减少了74.29%,运算速度提高了26.35%。综合网络模型的精度和参数分析,MDVNet可以在保证良好分割效率的基础上,实现高精度的腔镜手术场景分割。

3 分体式腔镜手术机器人

为了模拟腔镜手术机器人的实际操控情况,推进腔镜手术机器人的研究,基于Lap Game腹腔镜模拟器、Franka机械臂和内窥镜成像系统等设备构建了一套分体式腔镜手术机器人,并在该机器人上运行所提出的网络。团队搭建的腔镜手术机器人如图12所示,3个手术机械臂中两个机械臂搭载手术器械,一个机械臂搭载双目内窥镜。用户可以通过控制手柄操控搭载着手术器械的机器臂,在腹腔镜模拟器中完成切割、游离和缝合等腔镜手术操作,模拟腔镜手术过程。网络模型的

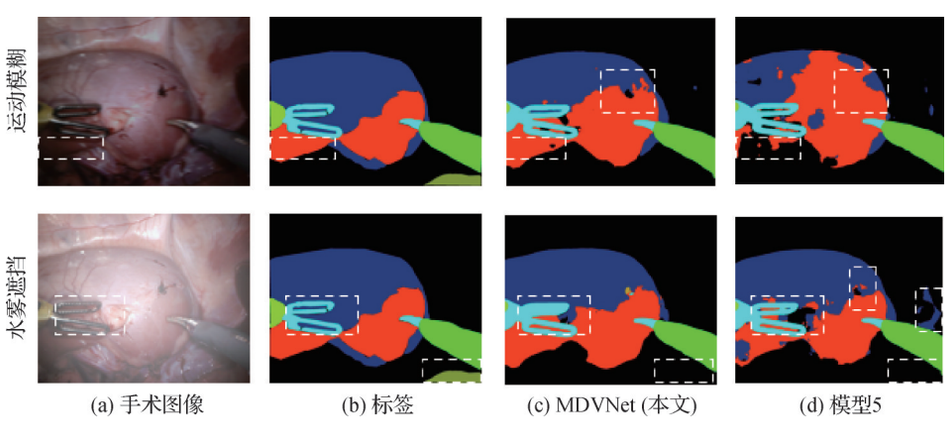


图11 在添加噪声的Endovis2018上的可视化结果

Fig. 11 Visualization results on Endovis2018 with added noises ((a) surgical images; (b) labels; (c) MDVNet(ours); (d) model 5)

表7 不同方法的参数

Table 7 Parameters of different methods

方法	参数量/M	FLOPs/G	FPS/(帧/s)
UNet(Ronneberger等,2015)	29.06	405.02	29.27
DeepLabv3+(Chen等,2018)	15.25	136.27	42.01
UPerNet(Xiao等,2018)	66.42	469.32	40.09
HRNetV2(Wang等,2021b)	65.86	186.62	13.33
OCRNet(Yuan等,2020)	70.38	322.34	10.47
STDCNet(Fan等,2021)	**12.61**	**23.48**	**53.42**
PSPNet(Zhao等,2017)	48.98	356.00	13.66
SegFormer(Xie等,2021)	24.73	35.83	17.41
SegNeXt(Guo等,2022)	27.57	63.91	15.28
LSKANet(Liu等,2024)	72.23	189.09	13.66
MDVNet(本文)	26.85	48.61	17.26

注:加粗字体表示各列最优结果。

图 12　腔镜手术机器人手术示例

Fig. 12　Operation of endoscopic surgical robot

分割结果会显示在用户控制端，辅助医生完成腔镜手术。

4　结　论

面向智能腔镜手术机器人的手术场景分割是指使用深度学习中的语义分割方法，分割腔镜手术场景中的解剖结构和手术器械等对象，并为每个像素分配类别标签。精确解析腔镜手术图像的整个场景是一项极具挑战性的任务，主要面临目标尺寸差异大、术中多源噪声和边界难以分辨 3 个挑战。基于此，本文提出了多尺度动态视觉网络 MDVNet。该网络采用编码器—解码器结构，在编码器部分，本文提出了动态大核卷积注意力模块 DLKA，通过多尺度大核注意力提取不同分割目标的多尺度特征，并通过动态选择机制进行自适应的特征融合，有效缓解了分割目标尺寸变化引起的网络误识别。在解码器部分，首先进行不同分辨率特征图的融合，通过低秩矩阵分解模块 LMD 滤除特征图中的噪声，获得高质量的特征图，并通过边界引导模块 BGM 引导模型学习手术场景的边界特征。MDVNet 在 3 个不同手术场景数据集上取得了 SOTA 结果，可以在团队搭建的分体式腔镜手术机器人顺利运行，通过将分割结果映射在手术视野中，辅助医生进行腔镜手术。本文方法的不足之处在于网络的推理速度有待进一步提高，未来腔镜手术场景分割的研究可以从以下几个方面展开：1）网络模型的推理速度：未来可以通过模型剪枝来移除模型中的冗余参数，大幅提高模型的推理速度，设计高效的实时分割模型，并部署在端侧设备上，为手术导航提供实时分割支持。2）无监督和半监督场景分割：标注高质量的手术场景分割数据需要大量的人力和医学专业知识，无监督和半监督学习技术能够利用大量未标注或少量标注的数据进行训练，从而大幅降低数据标注的成本和工作量，这对于推动该领域的发展至关重要。3）隐私和伦理问题：手术视频和图像数据涉及患者的隐私和医疗信息，后续在推进技术发展的同时，需要通过数据访问权限控制等方式保护患者隐私。在算法层面，也可以使用一些技术手段来保护隐私，如联邦学习、差分隐私和可解释性网络等，增强模型对隐私信息的防护能力。

致谢：本研究采用的数据得到了中南大学湘雅三医院的支持，在此表示感谢。

参考文献（References）

Allan M, Kondo S, Bodenstedt S, Leger S, Kadkhodamohammadi R, Luengo I, Fuentes F, Flouty E, Mohammed A, Pedersen M, Kori A, Alex V, Krishnamurthi G, Rauber D, Mendel R, Palm C, Bano S, Saibro G, Shih C S, Chiang H A, Zhuang J T, Yang J L, Iglovikov V, Dobrenkii A, Reddiboina M, Reddy A, Liu X T, Gao C, Unberath M, Kim M, Kim C, Kim C, Kim H, Lee G, Ullah I, Luna M, Park S H, Azizian M, Stoyanov D, Maier-Hein L and Speidel S. 2020. 2018 robotic scene segmentation challenge [EB/OL]. [2024-07-11]. https://arxiv.org/pdf/2001.11190.pdf

Chen L C, Zhu Y K, Papandreou G, Schroff F and Adam H. 2018. Encoder-decoder with atrous separable convolution for semantic

image segmentation//Proceedings of the 15th European Conference on Computer Vision. Munich, Germany: Springer: 833-851 [DOI: 10.1007/978-3-030-01234-2_49]

Collins T, Pizarro D, Gasparini S, Bourdel N, Chauvet P, Canis M, Calvet L and Bartoli A. 2021. Augmented reality guided laparoscopic surgery of the uterus. IEEE Transactions on Medical Imaging, 40(1): 371-380 [DOI: 10.1109/TMI.2020.3027442]

Despotović I, Goossens B and Philips W. 2015. MRI segmentation of the human brain: challenges, methods, and applications. Computational and Mathematical Methods in Medicine, 2015: #450341 [DOI: 10.1155/2015/450341]

D'Souza M, Gendreau J, Feng A, Kim L H, Ho A L and Veeravagu A. 2019. Robotic-assisted spine surgery: history, efficacy, cost, and future trends. Robotic Surgery: Research and Reviews, 6: 9-23 [DOI: 10.2147/RSRR.S190720]

Du H, Wang J Z, Liu M, Wang Y N and Meijering E. 2024. SwinPA-Net: swin transformer-based multiscale feature pyramid aggregation network for medical image segmentation. IEEE Transactions on Neural Networks and Learning Systems, 35(4): 5355-5366 [DOI: 10.1109/TNNLS.2022.3204090]

Fan M Y, Lai S Q, Huang J S, Wei X M, Chai Z H, Luo J F and Wei X L. 2021. Rethinking BiSeNet for real-time semantic segmentation//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 9711-9720 [DOI: 10.1109/CVPR46437.2021.00959]

Grammatikopoulou M, Flouty E, Kadkhodamohammadi A, Quellec G, Chow A, Nehme J, Luengo I and Stoyanov D. 2021. CaDIS: cataract dataset for surgical RGB-image segmentation. Medical Image Analysis, 71: #102053 [DOI: 10.1016/j.media.2021.102053]

Guo M H, Lu C Z, Hou Q B, Liu Z N, Cheng M M and Hu S M. 2022. SegNext: rethinking convolutional attention design for semantic segmentation//Proceedings of the 36th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc.: #84

Hendrycks D and Dietterich T. 2019. Benchmarking neural network robustness to common corruptions and perturbations//Proceedings of the 7th International Conference on Learning Representations. New Orleans, USA: ICLR: 1-16

Hua Y, Li Z Z, Pan J H and Yang X. 2023. Boundary-preserving multiscale glomerulus segmentation for full-stained kidney slice. Journal of Image and Graphics, 28(11): 3575-3589 (花勇, 李珍珍, 潘建宏, 杨烜. 2023. 边界信息保持的全染色肾脏切片多粒度分割. 中国图象图形学报, 28(11): 3575-3589) [DOI: 10.11834/jig.221025]

Jin Y M, Yu Y, Chen C, Zhao Z X, Heng P A and Stoyanov D. 2022. Exploring intra- and inter-video relation for surgical semantic scene segmentation. IEEE Transactions on Medical Imaging, 41(11): 2991-3002 [DOI: 10.1109/TMI.2022.3177077]

Lei B Y, Huang S, Li H, Li R, Bian C, Chou Y H, Qin J, Zhou P, Gong X H and Cheng J Z. 2020. Self-co-attention neural network for anatomy segmentation in whole breast ultrasound. Medical Image Analysis, 64: #101753 [DOI: 10.1016/j.media.2020.101753]

Liu D C, Li Q Y, Jiang T T, Wang Y Z, Miao R L, Shan F and Li Z Y. 2021. Towards unified surgical skill assessment//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 9517-9526 [DOI: 10.1109/CVPR46437.2021.00940]

Liu M, Han Y B, Wang J Z, Wang C, Wang Y N and Meijering E. 2024. LSKANet: long strip kernel attention network for robotic surgical scene segmentation. IEEE Transactions on Medical Imaging, 43(4): 1308-1322 [DOI: 10.1109/TMI.2023.3335406]

Lu C Y, Tang J H, Yan S C and Lin Z C. 2014. Generalized nonconvex nonsmooth low-rank minimization//Proceedings of 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus, USA: IEEE: 4130-4137 [DOI: 10.1109/CVPR.2014.526]

Ni Z L, Bian G B, Li Z, Zhou X H, Li R Q and Hou Z G. 2022. Space squeeze reasoning and low-rank bilinear feature fusion for surgical image segmentation. IEEE Journal of Biomedical and Health Informatics, 26(7): 3209-3217 [DOI: 10.1109/JBHI.2022.3154925]

Ronneberger O, Fischer P and Brox T. 2015. U-Net: convolutional networks for biomedical image segmentation//Proceedings of the 18th International Conference on Medical Image Computing and Computer-Assisted Intervention. Munich, Germany: Springer: 234-241 [DOI: 10.1007/978-3-319-24574-4_28]

Shen W T, Wang Y N, Liu M, Wang J Z, Ding R J, Zhang Z and Meijering E. 2023. Branch aggregation attention network for robotic surgical instrument segmentation. IEEE Transactions on Medical Imaging, 42(11): 3408-3419 [DOI: 10.1109/TMI.2023.3288127]

Su H, Mariani A, Ovur S E, Menciassi A, Ferrigno G and De Momi E 2021. Toward teaching by demonstration for robot-assisted minimally invasive surgery. IEEE Transactions on Automation Science and Engineering, 18(2): 484-494 [DOI: 10.1109/TASE.2020.3045655]

Wang J C, Jin Y M, Wang L S, Cai S T, Heng P A and Qin J. 2021a. Efficient global-local memory for real-time instrument segmentation of robotic surgical video//Proceedings of the 24th International Conference on Medical Image Computing and Computer Assisted Intervention. Strasbourg, France: Springer: 341-351 [DOI: 10.1007/978-3-030-87202-1_33]

Wang J D, Sun K, Cheng T H, Jiang B R, Deng C R, Zhao Y, Liu D, Mu Y D, Tan M K, Wang X G, Liu W Y and Xiao B. 2021b. Deep high-resolution representation learning for visual recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 43(10): 3349-3364 [DOI: 10.1109/TPAMI.2020.2983686]

Wang T M, Zhang X H, Zhang X B and Wang J C. 2019. Research progresses in laparoscopic augmented reality navigation. Robot, 41(1): 124-136 (王田苗, 张晓会, 张学斌, 王君臣. 2019. 腹腔镜增强现实导航的研究进展综述. 机器人, 41(1): 124-136) [DOI: 10.

13973/j.cnki.robot.170625］

Xiao T T, Liu Y C, Zhou B L, Jiang Y N and Sun J. 2018. Unified perceptual parsing for scene understanding//Proceedings of the 15th European Conference on Computer Vision. Munich, Germany: Springer: 432-448［DOI: 10.1007/978-3-030-01228-1_26］

Xie E Z, Wang W H, Yu Z D, Anandkumar A, Alvarez J M and Luo P 2021. SegFormer: simple and efficient design for semantic segmentation with transformers//Proceedings of the 35th International Conference on Advances in Neural Information Processing Systems. ［s. l.］: Curran Associates Inc.: #924

Xie Y L. 2022. Research on Laparoscopic Surgery Video Semantic Segmentation Algorithm Based on Deep Learning. Changsha: Hunan University (谢于廉. 2022. 基于深度学习的腔镜手术视频语义分割算法研究. 长沙: 湖南大学)［DOI: 10.27135/d.cnki.ghudu.2022.002634］

Xu W W, Xu L F, Li B K, Zhou X, Lyu N and Zhan S. 2024. TransAS-UNet: regional segmentation of breast cancer Swin Transformer and of UNet algorithm. Journal of Image and Graphics, 29(3): 741-754 (徐旺旺, 许良凤, 李博凯, 周曦, 律娜, 詹曙. 2024. TransAS-UNet: 融合 Swin Transformer 和 UNet 的乳腺癌区域分割. 中国图象图形学报, 29(3): 741-754)［DOI: 10.11834/jig.230130］

Yang L, Gu Y G, Bian G B and Liu Y H. 2023. A dual-encoder feature attention network for surgical instrument segmentation. Journal of Image and Graphics, 28(10): 3214-3230 (杨磊, 谷玉格, 边桂彬, 刘艳红. 2023. 双编码特征注意网络的手术器械分割. 中国图象图形学报, 28(10): 3214-3230)［DOI: 10.11834/jig.220716］

Yuan Y H, Chen X L and Wang J D. 2020. Object-contextual representations for semantic segmentation//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer: 173-190［DOI: 10.1007/978-3-030-58539-6_11］

Zhang X F, Zhang S, Zhang D H and Liu R. 2023. Group attention-based medical image segmentation model. Journal of Image and Graphics, 28(10): 3231-3242 (张学峰, 张胜, 张冬晖, 刘瑞. 2023. 引入分组注意力的医学图像分割模型. 中国图象图形学报, 28(10): 3231-3242)［DOI: 10.11834/jig.220748］

Zhao H S, Shi J P, Qi X J, Wang X G and Jia J Y. 2017. Pyramid scene parsing network//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 6230-6239［DOI: 10.1109/CVPR.2017.660］

作者简介

刘敏，男，教授，主要研究方向为机器人视觉感知、模式识别和图像处理。E-mail: liu_min@hnu.edu.cn

韩雨斌，通信作者，男，硕士研究生，主要研究方向为计算机视觉和医学图像处理。E-mail: hyb_hnu@163.com

秦敦璇，男，硕士研究生，主要研究方向为计算机视觉和医学图像处理。E-mail: qdx@hnu.edu.cn

陈祥，男，助理教授，主要研究方向为计算机视觉、医学图像分析、医学图像配准和三维网格重建。E-mail: xiangc@hnu.edu.cn

王耀南，男，中国工程院院士，主要研究方向为智能机器人技术、智能控制理论与应用、机器视觉感知与认知、智能自动化技术与系统和自主无人系统。E-mail: yaonan@hnu.edu.cn