目　录

摘要.....I

Abstract.....III

第一章 绪论.....1

1.1 研究背景和意义.....1

1.2 研究现状.....2

1.2.1 基于深度学习的辅助诊疗.....2

1.2.2 基于深度学习的医疗图像语义分割.....3

1.2.3 基于深度学习的医疗视频分类.....4

1.2.4 医疗信息管理系统.....5

1.2.5 存在问题.....6

1.3 研究内容和创新点.....6

1.4 论文结构.....8

第二章 基于改进 U-net 的共聚焦激光显微内镜杯状细胞语义分割方法.....11

2.1 引言.....11

2.2 数据集.....11

2.3 方法.....13

2.3.1 基于改进 U-net 的共聚焦激光显微内镜杯状细胞语义分割模型.....13

2.3.2 像素梯度特征提取模块.....14

2.3.3 损失函数.....16

2.3.4 算法描述.....17

2.4 实验和讨论.....17

2.4.1 评价指标.....17

2.4.2 实验细节.....18

2.4.3 实验结果 19
2.4.4 讨论 20
2.5 本章小结 22
第三章 结合内窥镜视频的共聚焦激光显微内镜检测部位识别方法 23
3.1 引言 23
3.2 材料 25
3.3 方法 25
3.3.1 方法概述 25
3.3.2 共聚焦激光显微内镜检测部位识别模型 26
3.3.3 损失函数 31
3.3.4 算法描述 32
3.4 实验和讨论 33
3.4.1 实验细节 33
3.4.2 实验结果 33
3.4.3 讨论 34
3.5 本章小结 36
第四章 面向共聚焦激光显微内镜的临床诊断数据管理系统 38
4.1 引言 38
4.2 需求分析 38
4.3 系统设计 39
4.3.1 系统架构 39
4.3.2 功能模块设计 40
4.3.3 数据管理 40
4.3.4 杯状细胞语义分割 41
4.3.5 数据库设计 42

4.4 系统实现 43
4.4.1 系统框架 43
4.4.2 开发工具 44
4.4.3 功能实现 44
4.5 系统测试 50
4.6 本章小结 53
第五章 总结与展望 54
5.1 总结 54
5.2 展望 55
参考文献 56
攻读硕士学位期间的主要成果 64
致　谢 65

摘要

胃癌严重威胁着人类的生命健康，已经成为第三大癌症致死率病因以及第五大新增病例。研究证明，早期发现及治疗是减少胃癌发病率，提高患者生存率的关键。其中，胃黏膜肠上皮化生（Gastric Intestinal Metaplasia，GIM）是 Correa 序列中胃癌发生的中间步骤，被视为是重要的癌前病变。早先胃黏膜肠上皮化生的诊断主要依赖于观察染色后的组织结构，缺乏细胞层次的微观数据。随着数字影像近十年的飞速发展，研究者将共聚焦显微技术应用到临床诊疗中，研发了共聚焦激光显微内镜（Confocal Laser Endomicroscopy，CLE），能够从微观层面反应病情严重程度。作为一种新型诊断工具，共聚焦激光显微内镜能更为细致地分析病情变化，做出更为精确的病情评估。

在胃黏膜肠上皮化生领域，共聚焦激光显微内镜临床诊断时间长，同区域下包含信息更为丰富，给医生临床诊断带来了困难。同时，受限于成像技术，仅通过共聚焦激光显微内镜无法识别检测部位，进一步加大了漏诊的风险。为了解决上述问题，论文将人工智能技术引入到共聚焦激光显微内镜临床诊疗中，提出了基于改进 U-net 的杯状细胞语义分割方法、共聚焦激光显微内镜检测部位识别方法并研发了临床数据管理系统。主要研究内容如下：

（1）针对临床共聚焦激光显微内镜图像中杯状细胞区域难以分割的问题，提出了基于改进 U-net 的杯状细胞语义分割方法。方法将特征提取融入 U-net 模型的同层链接，并利用得到的特征图引导模型上采样。最后，方法基于模型的输出计算概率图，并生成最终的分割结果。实验证明，改进的 U-net 能够进一步过滤浅层特征区域，并引导网络深层特征，有效提高杯状细胞分割准确率。

（2）针对共聚焦激光显微内镜检测过程中无法识别被检测部位的问题，提出了结合内窥镜视频的共聚焦激光显微内镜检测部位识别方法。为了实现在共聚焦激光显微内镜检测过程中的部位特征提取，提出了一种基于沙漏卷积和时序特征的共聚焦激光显微内镜检测部位识别模型。即采用沙漏卷积提取单帧关键特征、偏移模块以及挤压机制提取时序特征，并通过全连接神经网络预测检测部位。实验证明，提出的方法能有效地提取关键特征，精准识别检测部位。

（3）针对共聚焦激光显微内镜数据管理困难的问题，研发了共聚焦激光显微内镜临床数据管理系统。通过采用前后端分离的开发模式，实现了临床视频和图像的管理、杯状细胞区域在线分割等功能。

关键字：胃黏膜肠上皮化生；共聚焦激光显微内镜；部位识别；语义分割；注意力机制。

Abstract

Gastric cancer poses a serious threat to human life and health, ranking as the third leading cause of cancer-related deaths and the fifth leading new case. Research has shown that early detection and treatment are crucial for reducing the incidence of gastric cancer and improving patient survival rates. Among them, gastric intestinal metaplasia is an intermediate step in the correa sequence of gastric cancer development and is considered an important precursor lesion. Previously, the diagnosis of gastric intestinal metaplasia mainly relied on observing tissue structures after staining, lacking microscopic data at the cellular level. With the rapid development of digital imaging technology in the past decade, researchers have applied confocal microscopy technology to clinical diagnosis, developing confocal laser endomicroscopy, which can reflect the severity of the condition at the microscopic level. As a new diagnostic tool, confocal laser endomicroscopy can more meticulously analyze illnesses and make more accurate assessments.

In gastric intestinal metaplasia diagnosis, confocal laser endomicroscopy (CLE) has a long clinical diagnosis time and contains richer information within the same area, which poses challenges for clinical diagnosis by doctors. At the same time, limited by imaging technology, the detected areas cannot be directly obtained only through CLE. To address these issues, this paper introduces artificial intelligence technology into clinical diagnosis and treatment in CLE, proposing a goblet cell segmentation method based on improved U-net, a CLE diagnosis area recognition method, and developing a clinical data management system. The main research contents are as follows:

(1) To address the challenge of goblet cell segmentation in confocal laser endomicroscopy, a goblet cell segmentation method based on an improved U-net is proposed. The method integrates feature extraction into the homo-layer connections of the U-net model and utilizes the obtained feature maps to guide model upsampling. Finally, the method calculates probability maps based on the model's output and generates the final segmentation results. Experimental results demonstrate that the improved U-net can further filter shallow feature regions, guide deep network features, and effectively improve the accuracy of goblet cell segmentation.

(2) To address the issue of the diagnosis area identification during confocal laser endomicroscopy (CLE), a method combining endoscopic video with CLE for area identification is proposed. To extract area features during CLE, an area identification model based on hourglass convolution and temporal features is introduced. The hourglass convolution is utilized to extract key features from single frames, offset modules and attention mechanisms are utilized to extract temporal features, and predict the detected areas through a fully connected neural network.

Experimental results demonstrate that the proposed method effectively extracts key features and accurately identifies the diagnosis area.

(3) To address the issue of management difficulties of confocal laser microscopy data, a clinical data management platform for confocal laser endomicroscopy was developed. By adopting a front-end and back-end separation development mode, we have accomplished the management of clinical videos and images, as well as online the goblet cell segmentation method.

Keywords: Gastric Intestinal Metaplasia; Confocal Laser Endomicroscopy; Part Recognition; Semantic Segmentation; Attention Mechanism。

第一章 绪论

1.1 研究背景和意义

胃癌作为第三大癌症致死率病因以及第五大新增病例[1-3]，严重威胁着人类的生命健康。医学研究表明，早期发现及干预是减少胃癌发病率、提高患者生存率的关键。Correa等人[4]的研究证明，胃癌是由慢性胃炎、萎缩性胃炎、肠上皮化生、上皮内瘤变发展而来的，这一系列变化在胃癌发生学中称为 Correa 序列。其中，胃黏膜肠上皮化生（Gastric Intestinal Metaplasia，GIM）作为 Correa 序列中胃癌发生多步骤假设的中间步骤，已被视为癌前病变[5]。化生的肠上皮是一种高度分化的上皮，最明显的特征是胃粘膜中出现肠化细胞。临床中，胃黏膜中出现的肠化细胞由于外形类似于杯状，也被称为杯状细胞（Goblet Cells，GC）。受到医疗影像发展的影响，早先胃黏膜肠上皮化生的诊断主要依赖于观察染色后的组织结构。然而，此种方法缺乏细胞结构（Cellular Structure，CS）的微观数据，无法做到细胞层次的病情量化分析。

得益数字影像近十年的飞速发展，研究者将共聚焦显微成像技术应用到临床诊疗中，发明出共聚焦激光显微内镜（Confocal Laser Endomicroscopy，CLE）。共聚焦激光显微内镜作为一种新型技术，可实时观察放大 1000 倍的发病部位，能够从微观层面反应病情严重程度。共聚焦激光显微内镜作为一种诊断工具，能反映出更为细致的病情变化，做出更为精确的病情评估，使得从细胞层次诊断患者病情成为可能[6-7]。近年来，共聚焦激光显微内镜已推动了诸多医学领域的发展。

相对于宏观结构，微观结构（细胞层面）能够更准确地反应病情细微变化。这有助于提供更为确切的诊断，便于进行更为精细化的治疗。然而，细胞层面的观察数据量更大，信息更为丰富，这也就造成了诊断上的困难。仅依赖人工识别杯状细胞等病变区域，难免会出现遗漏的情况。另一方面，微观结构无法识别检测部位，这使得医生在诊断时需同时关注内窥镜和共聚焦激光显微内镜以确定检测部位。进一步增加了误诊、漏诊的风险。

为克服上述问题，本文将深度学习（Deep Learning，DL）技术引入共聚焦激光显微

内镜检测胃黏膜肠上皮化生领域，分别在杯状细胞语义分割领域、共聚焦激光显微内镜检测部位识别领域进行了基于人工智能辅助诊疗方向的研究。最后，为了便于临床数据的管理，研发了针对共聚焦激光显微内镜临床数据的信息管理系统。该系统在提供数据管理的同时，将杯状细胞语义分割模型融入其中，有效提高了临床数据管理质量。

1.2 研究现状

1.2.1 基于深度学习的辅助诊疗

近年来，基于深度学习的辅助诊疗技术取得了长足进步[8-10]，在医学图像识别[11]、电子健康记录挖掘[12]、慢性病管理[13]等领域取得了重大进展。通过深度神经网络模型评估病情严重程度、干预诊断过程，能起到提高诊断效率、降低误诊率的作用。其中，Nie 等人[14]提出了一种使用多通道数据的 3D CNN 结构用于提取磁共振成像的高级别胶质瘤特征，并通过训练支持向量机来预测患者生存时间。Jnawali 等人[15]首先将 3D CNN 模型用于大型横截面医学图像数据分析，并提出了根据横截面 CT 图像检测脑出血的智能方法。Milletar 等人[16]探索了网络模型在磁共振成像中经颅超声（Transcranial Ultrasound，TU）体积分割中的应用。此外，Dou 等人[17]提出了两级全连接 3D CNN 架构以从磁共振加权图像（Susceptibility-weighted Imaging，SWI）中识别脑微出血（Brain Microbleeds，BM）区域，有效排除了许多假阳性病人，降低了患者误诊概率。

在胃疾病辅助诊疗领域，诊断过程中的胃部位（Stomach Areas，SA）识别已取得了诸多进展[18]。实时监测检查部位能有效控制检测时间、提高内镜检测质量。其中，Wu 等人[19]将胃的解剖位置分为 10 个粗分部位和 26 个细分部位，并采用改进的深度卷积神经网络（Deep Convolution Neural Network，DCNN）进行解剖分类。最终准确率分别为 90%和 65.9%，有效减少了胃癌内窥镜的检测盲点。He 等人[20]将内窥镜解剖结构划分为 11 个部位，使用 DenseNet121[21]的分类准确率达到 91.11%，为临床肠镜检查提供了更为详细的分析数据。相较于上述研究，共聚焦激光显微内镜检测过程中的内窥镜成像探针与胃壁距离更近，视野范围更为狭窄，气泡、胃蠕动、粘液对成像的影响更大。这使得单一图像的部位识别更加困难且不可靠。为此，目前尚未有有效的共聚焦激光显微内镜部位识别模型被提出。

1.2.2 基于深度学习的医疗图像语义分割

在图像语义分割（Image Semantic Segmentation，ISS）方向，卷积结构凭借着对颜色、纹理特征的捕获能力在医学图像处理领域占据着重要地位。起初，基于卷积的神经网络[22-23]使用特定像素周围的一个图像块作为输入，通过分类函数获得对应像素的条件概率。然而，此类方法需要大量的计算资源，不够灵活。Jonathan 等人[24]提出了全卷积神经网络，通过引入下采样操作，降低了模型训练成本，提高了模型训练效率。然而，此网络在将深层特征信息通过线性上采样恢复原始尺寸过程中会损失边界特征信息。为此，Ronneberger 等人[25]通过改进全卷积神经网络提出了 U-net 结构。通过同层连接以及上采样过程中的卷积，使得模型在上采样阶段仍能对边界特征进行一定程度上的特征提取，大大提高了边界像素分类的精准度。在后续的研究中，Valanarasu 等人[26]对 U-net 模型进行了进一步改进，提出了风筝网络结构，即通过反卷积与卷积的融合，进一步提高了模型的边界信息捕获能力。

然而，卷积结构由于受到感受野的限制，无法捕获目标的整体结构特征[27]。这使得网络难以关注到目标的形状、大小等信息。同时。识别到的区域会存在尾影（识别边界存在部分点状、块状区域）、孤岛（目标内部区域未能准确识别）区域，严重降低了模型识别准确率。为此，一些研究通过将卷积结构与机器学习相结合，优化网络识别区域。其中，一种策略是通过线性流程的方式，在利用深度网络对像素进行分类后，再使用机器学习对区域目标进行分类[28]。另一种策略是通过改进概率模型优化识别区域。郑帅等人[29]将条件随机场中的马尔可夫概率转化为循环神经网络，实现了端对端的训练。他们提出的模型将卷积网络识别到的信息与利用条件随机场得到的概率进行线性相加，实现了对网络参数的训练。另一些研究通过改进深度神经网络的部分结构，并使用改进后的损失函数引导特征提取过程，从而实现对识别区域的优化[30-32]。总的来说，目前大多数针对卷积神经网络的改进主要集中在针对纹理、边界等特征，而对目标区域的整体特征（如形状）提取上仍存在一定不足。

伴随着深度学习的发展，自注意力机制的提出能有效解决卷积结构无法获得长距离依赖的问题。Vaswan 等人[33]首先提出了理论上可获得全局特征的自注意力机制，并应用到自然语言领域。在此基础上，Dosovitskiy 等人[34]提出的 vision transformer 将这一结构应用到图像处理领域。vision transformer 通过将图像的切割与展开，实现了自注意力机制对

图像的特征提取。陈洁能等人[35]也将这一结构应用到医学图像语义分割任务中，提出了Transunet 模型。Transunet 利用卷积与自注意力机制的融合，有效增强了模型全局特征提取能力，提高了图像语义分割的准确率。在后续的研究中，卷积与注意力机制结合的特征提取方式取得了一定成果。然而，此类结构往往需要大量的训练数据，这使得训练成本非常昂贵。为此，Valanarasu 等人[36]提出了门控注意力机制。门控注意力机制通过控制模型的输入降低学习成本，但此种结构也使得网络在一定程度上丧失了长距离判断能力，丧失了自注意力机制的独特优势。

1.2.3 基于深度学习的医疗视频分类

在视频分类（Video Classification，VC）任务上，最初的特征提取主要是基于手工选择来完成。其中，改进轨迹的动作识别[37]（Improved Dense Trajectories，iDT）被认为是最有效的结构之一。伴随着深度学习的发展，端到端的深度网络模型逐步应用到视频分类任务中。Zha 等人[38]提出了用于视频分类的 2DCNN 模型。模型通过 2DCNN 提取帧级特征，并通过一些现有的分类模型（例如 SVM 等）预测视频类别。另一方面，Vosta 等人[39]提出了 2.5D 的视频特征提取方式，并构建了 RNN-CNN 网络模型。此类模型通过 2DCNN 等结构提取单帧图像特征，并通过循环神经网络提取时序特征，从而实现了视频特征的提取。上述模型避免了人工选择特征所带来的繁琐，提高了视频分类的准确率。在帧级提取特征的 2D CNN 模型取得成功后，3DCNN 也逐渐应用到视频特征提取过程中。C3D[40]将视频特征通过 3D 卷积进行提取，并通过全连接神经网络（Fully Connected Neural Network，FCNN）实现了视频的精准分类。

在特征提取结构的改进中，光流的引入被认为是视频分类中最成功的应用之一。TSN[41]将光流引入到视频分类任务中，并提出了双流视频分类网络。通过引入光流场，双流视频分类网络显著提高了现有模型的分类准确率，并实现了在少量视频数据训练下的精准分类。然而，光流的引入也造成了长视频中时序信息的不敏感，以及容易过拟合等问题。另一方面，TSM[42]偏移模块的引入也被认为是最有效的时序特征学习方式之一，通过时序维度的偏移，模型在仅少量增加计算量的同时，提高了对时序信息的学习能力。

伴随着 Transformer 的发展，自注意力机制也逐渐应用到视频分类任务中。Liu 等人[43]将视频以 3D 切割的方式最先将自注意力机制应用到视频分类任务中。Bertasius[44]提出了

TimeSformer 的视频理解架构。TimeSformer 结构完全基于 Transformer，将输入视频视为从每一帧中提取的图像块（patches），通过时空自注意力机制提取输入序列的时序特征。然而，此类模型往往需要大量的训练样本，这必然带来数据集成本的上升。这为自注意力机制在医学研究中的普及带来了不小的困难。

1.2.4 医疗信息管理系统

近年来，伴随着人工智能的发展，越来越多的生产领域通过引入人工智能来提高生产效率。其中，医疗领域作为最成功的人工智能试点之一，引起国内外研究机构、企业的广泛关注[45-46]。全球范围内的各大医疗机构，如梅奥、克里夫兰等[47-48]，正在积极寻求与人工智能公司合作，旨在辅助疾病的探测、诊断、治疗和管理[49]。在人工智能落地医疗生产过程中，训练数据集的构建以及管理作为模型落地的关键步骤，影响着后续的识别结果、临床应用及决策表现等。

起初，临床数据管理主要依赖于文件系统和表格记录等方式。然而，这种方式不仅操作繁琐，而且导致了数据难以回溯。医疗信息管理系统的提出对于现代医疗体系具有深远的意义[50-51]。首先，医疗信息管理系统为医疗机构提供了高效的数据管理和存储手段，显著提升了医疗服务的质量和效率。通过电子化的患者信息管理，医生可以更便捷地获取患者病历资料，实时了解病情发展趋势，有助于更快速、准确地制定诊疗方案[52]。此外，医疗信息管理系统为医学研究和医疗决策提供了宝贵的数据支持。通过对大量患者数据的分析，可以发现疾病的潜在规律，制定更有效的治疗方案，并为公共卫生政策提供科学依据，进而促进医学科研和医疗实践的不断进步[53]。最重要的是，系统有助于增强患者与医护之间的沟通与信任。患者可以方便地查阅自己的病历信息、预约医生、了解诊疗进展，从而更积极地参与自己的健康管理[54]。这种信息透明性不仅提升了患者满意度，也有助于建立更紧密的医患关系。

综上所述，医疗信息管理系统的建设和运用不仅是医疗服务数字化的必然趋势，更是推动医疗行业升级、提高服务水平、优化资源配置的重要举措。近年来，Epic Systems[55]、Cerner Millennium[56]、McKesson Paragon[57]等医疗系统的应用，加快了医疗信息的协同共享，促进了医疗团队更好地协同工作，在支持医生制定治疗计划、提高医疗流程效率方面发挥了积极作用。

1.2.5 存在问题

随着计算机和人工智能技术的进步，利用人工智能辅助临床诊疗受到了广泛关注。在共聚焦激光显微内镜诊断胃黏膜肠上皮化生领域，目前主要存在以下问题：

（1）共聚焦激光显微内镜中杯状细胞的分布区域识别是诊断胃黏膜肠上皮化生严重程度的必要环节。然而，手动分割耗时、繁琐，且受主观因素影响较大。同时，由于共聚焦激光显微内镜中存在的干扰区域以及成像分辨率较大，现有的分割方法性能较差。

（2）共聚焦激光显微内镜检测是识别胃黏膜肠化生严重程度的重要检测手段。临床采用的评估方式需要全面检测胃窦、胃体和胃角。然而，微观成像无法直接判断检测部位，存在遗漏潜在疾病发生部位的风险。现有深度神经网络由于受到检测视频中干扰帧的影响，无法利用内窥镜视频准确识别出检测部位。

（3）共聚焦激光显微内镜诊断通常持续 10-20 分钟，期间产生的临床数据目前常采用文件系统进行存储。此种存储方式使得临床数据管理混乱，造成了数据管理效率低、数据回溯困难等问题。

1.3 研究内容和创新点

围绕共聚焦激光显微内镜在临床上面临的杯状细胞区域难以分割、检测区域难以确定、数据管理困难的问题，论文主要做了以下研究：

（1）共聚焦激光显微内镜杯状细胞语义分割方法

共聚焦激光显微内镜中，杯状细胞的分布区域识别对诊断肠上皮化生至关重要。然而，手工分割费时费力，过于繁琐、主观。另一方面，由于受到干扰区域、成像分辨率的影响，现有深度神经网络的分割性能较差。为此，本文提出了一种基于改进 U-net 的共聚焦激光显微内镜杯状细胞语义分割方法，并提出了基于改进 U-net 的共聚焦激光显微内镜杯状细胞语义分割模型。模型将像素梯度注意力机制融合到同层链接中，用于学习杯状细胞周围的像素梯度信息。通过关注像素梯度变化，避免模型将染色液泄露、正常细胞等区域识别为杯状细胞，进而提高模型分割准确率。为实现该方法，来自齐鲁医院 60 位临床病例的 343 张共聚焦激光显微内镜图片被收集用于模型的训练。收集的图片由经验丰富的医师进行杯状细胞区域标注。标注后的图片经过相关领域专家的反复校对以保证准确性。

在杯状细胞语义分割数据集上，引入像素梯度注意力机制的 U-net 模型在测试数据集上 IOU 达到了 87.95%，DICE 达到了 86.64%，结果优于目前所提出的大多数算法。实验结果显示，基于深度学习的模型在共聚焦激光显微内镜图像处理上表现出了巨大潜力。实验证明，临床中人工针对共聚焦激光显微内镜图像的处理与深度网络模型表现所差无几。这项研究表明，通过引入人工智能能快速帮助临床医师做出判断，保障诊断结果的准确率，有效降低人工成本。这项工作在利用共聚焦激光显微内镜评估肠化生严重程度、取代活检检测肠化生等研究方向上有着巨大的应用潜力。

（2）共聚焦激光显微内镜诊断部位识别方法

共聚焦激光显微内镜检测是识别胃黏膜肠化生严重程度的重要检测手段。目前临床采用的评估方式是医生通过分析胃窦、胃体和胃角部位的共聚焦激光显微内镜视频信息帧片段，综合分析病人肠化生严重程度。然而，由于微观成像结构的限制，临床上无法通过共聚焦视频识别出被检测部位。存在遗漏潜在疾病发生部位的风险，不利于后续对发病部位的精准治疗。为此，本文提出了结合内窥镜视频的共聚焦激光显微内镜检测部位识别方法。首先，通过实时检测共聚焦激光显微内镜工作状态，获得信息帧序列起始时间。其次，通过共聚焦激光显微内镜视频与内窥镜视频在时间序列上的对应关系，提取信息帧片段前 1.6 秒的内窥镜视频信息。最后，本文设计了基于沙漏卷积的单帧关键特征提取模块和时序敏感的空间特征提取模块的共聚焦激光显微内镜检测部位识别模型，用于识别共聚焦激光显微内镜检测部位。

为验证所提出的方法，来自齐鲁医院的 67 段临床共聚焦激光检测视频被收集用于网络训练。每段视频包含一段内窥镜视频以及一段共聚焦激光显微内镜视频。根据共聚焦激光显微内镜与内窥镜在时序上的对应关系，提取了 500 段共聚焦激光显微内镜起始信息帧对应的前 1.6 秒内窥镜视频信息。收集到的数据经过人工标注，构建了共聚焦激光显微内镜部位检测数据集。数据集包含内窥镜视频以及对应的部位标签。共聚焦激光显微内镜检测部位识别模型在测试集上的准确率达到了 97.1%，实现了共聚焦激光显微内镜检测部位的精准识别。这项研究能有效降低部位漏诊率，确保检查的全面性，提高检测质量，减轻医生的工作负担。

（3）面向共聚焦激光显微内镜的临床诊断数据管理系统

本文研发了一种针对共聚焦激光显微内镜的数据管理系统，系统针对共聚焦激光显

微内镜所产生的临床数据，管理图像、视频类型的数据信息。系统采用前后端分离的设计方式。前端通过 Vue 进行搭建，后台采用 Django 框架进行管理，中间层通过 Axios 连接。受限于数据存储量大的问题，系统采用文件系统结合数据库的数据存储方式，在服务器端存储临床数据。用户可将数据上传至服务器中进行存储，并通过 Web 浏览器进行访问。针对共聚焦激光内镜杯状细胞难以识别的问题，系统还提供了杯状细胞语义分割功能。具体而言，系统通过 Flask 框架，将基于改进 U-net 的共聚焦激光显微内镜杯状细胞语义分割模型部署到系统。用户可将本地临床图像上传至系统中。系统通过网络模型识别杯状细胞区域，并将识别到的结果提供给用户。

论文的创新点主要包括：

（1）针对临床共聚焦激光显微内镜图像中杯状细胞区域难以分割的问题，提出了基于改进 U-net 的杯状细胞语义分割方法。方法将特征提取融入 U-net 模型的同层链接，并利用得到的特征图引导模型上采样。最后，方法基于模型的输出计算概率图，并生成最终的分割结果。实验证明，改进的 U-net 能够进一步过滤浅层特征区域，并引导网络深层特征，有效提高杯状细胞语义分割准确率。

（2）针对共聚焦激光显微内镜检测过程中无法识别被检测部位的问题，提出了结合内窥镜视频的共聚焦激光显微内镜检测部位识别方法。为了实现在共聚焦激光显微内镜检测过程中的部位特征提取，提出了一种基于沙漏卷积和时序特征的共聚焦激光显微内镜检测部位识别模型。即采用沙漏卷积提取单帧关键特征、偏移模块以及挤压机制提取时序特征，并通过全连接神经网络预测检测部位。实验证明，提出的方法能有效地提取关键特征，精准识别检测部位。

（3）针对共聚焦激光显微内镜数据管理困难的问题，研发了共聚焦激光显微内镜临床数据管理系统。通过采用前后端分离的开发模式，实现了临床视频和图像的管理、杯状细胞区域在线分割等功能。

1.4 论文结构

本文针对共聚焦激光显微内镜在早期胃癌临床诊断领域存在的问题展开研究，提出了基于改进 U-net 的共聚焦激光显微内镜杯状细胞语义分割方法、结合内窥镜视频的共聚焦激光显微内镜检测部位识别方法。进一步的，针对临床数据难以管理的问题，本文还

研发了共聚焦激光显微内镜临床数据管理系统，并将杯状细胞语义分割模型部署到系统中。

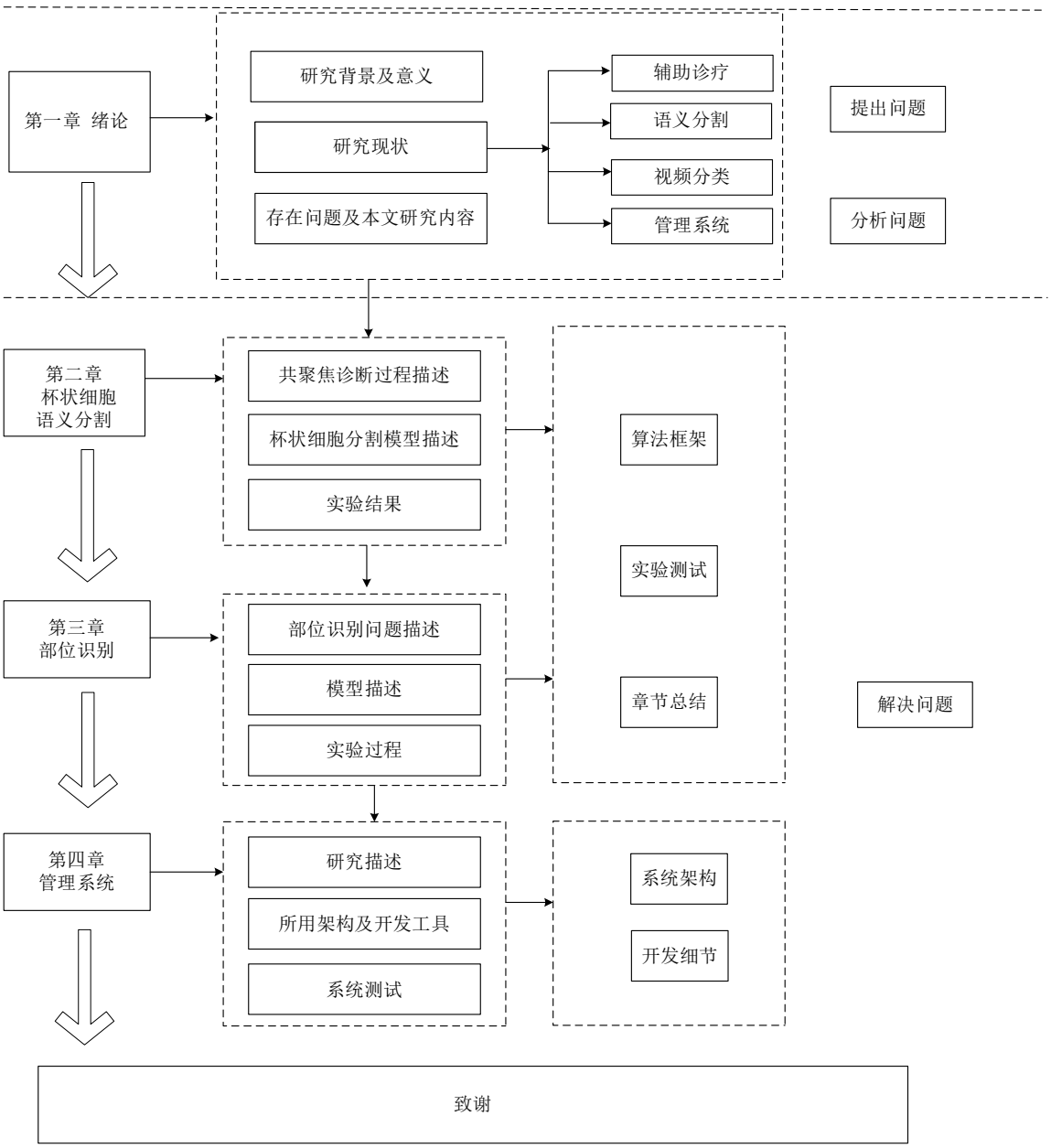

图 1-1 论文结构

论文的内容结构图如 1-1 所示，各章研究内容如下：

第一章：课题研究意义及背景。对共聚焦激光显微内镜以及早期胃癌诊断的研究背景、相关工作进行了简单介绍，最后阐明文章的结构及主要研究工作。

第二章：针对杯状细胞语义分割进行了详细描述。首先介绍了引入像素梯度注意力机制的 U-net 模型。在此基础上，列举出了相关实验结果以证明引入像素梯度注意力机制 U-net 模型的合理性以及优越性。

第三章：本章介绍了临床上共聚焦激光显微内镜诊断部位难以确定的问题，并提出了结合内窥镜视频的共聚焦激光显微内镜检测部位识别方法。针对此方法，本章还介绍了在共聚焦激光显微内镜部位检测数据集上进行的一系列实验，并验证了所提方法的可靠性。

第四章：本章针对临床信息难以管理的问题，介绍了面向共聚焦激光显微内镜临床诊断数据管理系统。该系统提供了针对视频以及图像两种数据的管理方案，并部署了杯状细胞语义分割模型。

第五章：总结本文主要研究内容和创新点，并对下一步工作进行展望。

第二章 基于改进 U-net 的共聚焦激光显微内镜杯状细胞语义分割方法

2.1 引言

杯状细胞分布作为癌变前的重要特征，在临床诊断、肠化生治疗以及术后检查中具有重要参考价值。通过分析杯状细胞的分布，能有效判断胃黏膜肠上皮化生严重程度，具备较高的特异性。临床上，杯状细胞分布区域识别作为病情诊断的重要环节，受到了广泛关注。另一方面，共聚焦激光显微内镜图像分辨率较高，所需观测范围更广，人工观察费时费力，诊断评估也更为困难。为此，本章提出了基于改进 U-net 的共聚焦激光显微内镜杯状细胞语义分割方法。通过该方法辅助医生诊断，可起到降低人力成本、提高诊断效率、降低漏诊率的作用。

基于改进 U-net 的共聚焦激光显微内镜杯状细胞语义分割方法（Goblet Cell Segmentation Method from Confocal Laser Endomicroscopy with an improved U-Net，GCSCLE）选取卷积作为主干特征提取结构，通过下采样捕获特征在高维上的关系，并缓解模型训练压力。通过上采样利用深层特征预测识别结果。为了使模型关注到颜色梯度特征以及形状、大小信息，GCSCLE 改进了像素注意力机制[58]，提出像素梯度特征提取模块。具体而言，GCSCLE 将浅层特征通过像素梯度特征提取模块提取像素梯度特征，并利用提取到的特征引导上采样关注区域，进而使模型关注到像素梯度变化。实验证明，模型通过捕获像素梯度信息，能有效关注识别区域的形状、大小以及边界梯度，进而有效区分杯状细胞区域与正常细胞区域。分割结果表明，模型在共聚焦激光显微内镜杯状细胞语义分割任务中的表现接近人工标注，临床中能有效辅助医生诊断，减轻病理科医生的工作负担。

2.2 数据集

杯状细胞语义分割数据集纳入了 62 名受试者的 334 张共聚焦激光显微内镜临床图像以组成训练集与测试集。32 人为男性，30 人为女性。其中，测试集包含 12 名受试者的 80 张临床图像，训练集包含 254 张。这些受试者均患有不同程度的肠化生，但这些患者可能

同时患有其他疾病。经验丰富的内镜科医师使用共聚焦激光显微内镜，获取胃部不同部位的临床图像。病理科医师检查图像中的每个像素信息，并将其中所属杯状细胞区域的像素进行手工标注。图 2-1 显示了杯状细胞标注示意图。特别的，背景类（包含纤毛、细胞组织液区域、正常细胞区域）必须包含在模型的输出中，但未标明。

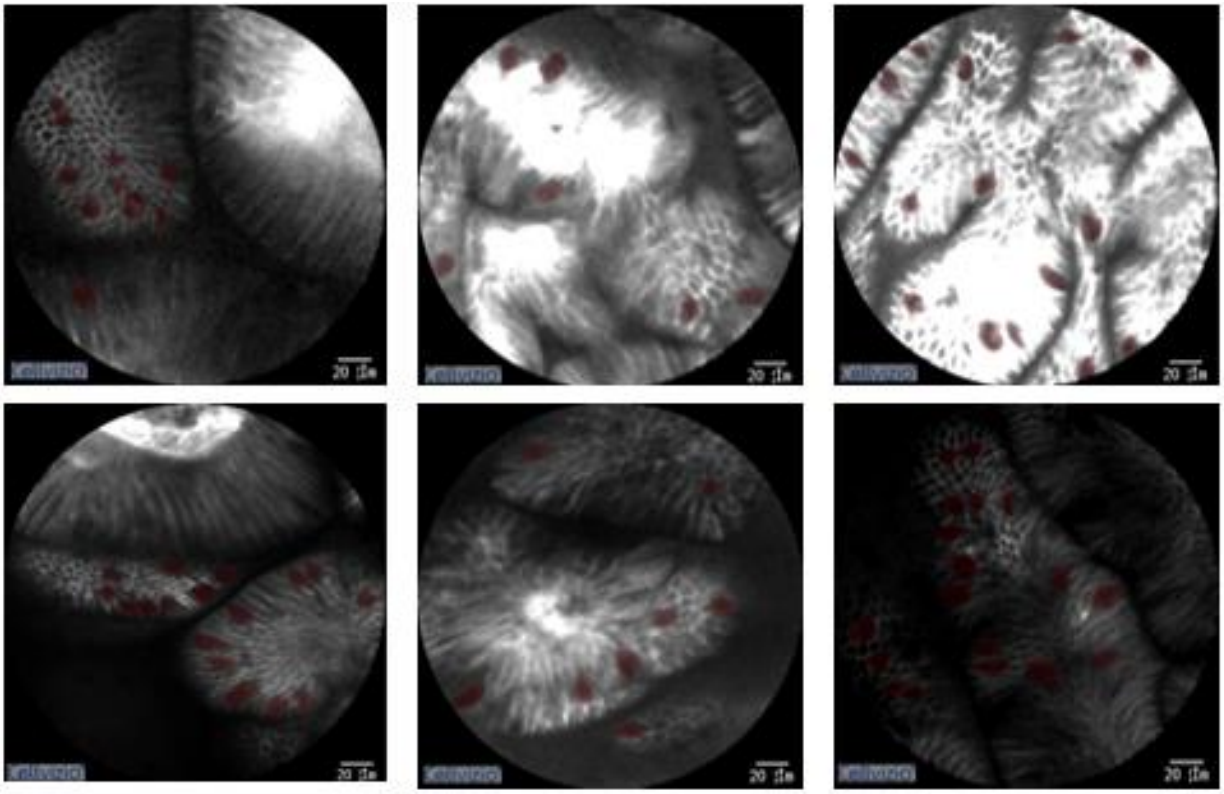

图 2-1 共聚焦激光显微内镜标注效果图

共聚焦激光显微内镜的真值标注由病理科医师根据相关研究[59]完成。通过多位医师的共同标注，反复校对，最终得到标注图像。此项工作由经验丰富的 10 名病理科医师参与。为提高标注效率，标注医生使用开源软件 labelme 进行标注。每张图像共进行了三轮数据标注。在第一轮的标注中，每张图像安排三位不同医师进行数据标注。通过对比标注信息，找出含有歧义的区域。此后，再通过第二轮次的校对修正，以及第三轮次的专家检验，完善标注信息。在标注中，杯状细胞特征的主要依据有以下三点：

（1） 颜色特征：经过染色后的杯状细胞成像往往为深灰色区域。

（2） 大小特征：相较于正常细胞，杯状细胞往往较大。

（3） 颜色梯度特征：细胞区域周围存在渐变颜色区域信息，此特征区域可以排除因细胞质泄露而导致的灰黑色区域。

2.3 方法

基于改进 U-net 的共聚焦激光显微内镜杯状细胞语义分割方法针对共聚焦激光显微内镜临床图像，实时获取图像中的杯状细胞区域。首先，方法通过共聚焦激光显微内镜检测，获取临床图像。其中，图像的获取可通过医生主动抓拍以及实时监听。其次，利用基于改进 U-net 的共聚焦激光显微内镜杯状细胞语义分割模型，获取输入图像对应的像素分类概率图（模型参数由调参、训练所得）。最后，通过得到的条件概率图，预测图像中的杯状细胞区域。

2.3.1 基于改进 U-net 的共聚焦激光显微内镜杯状细胞语义分割模型

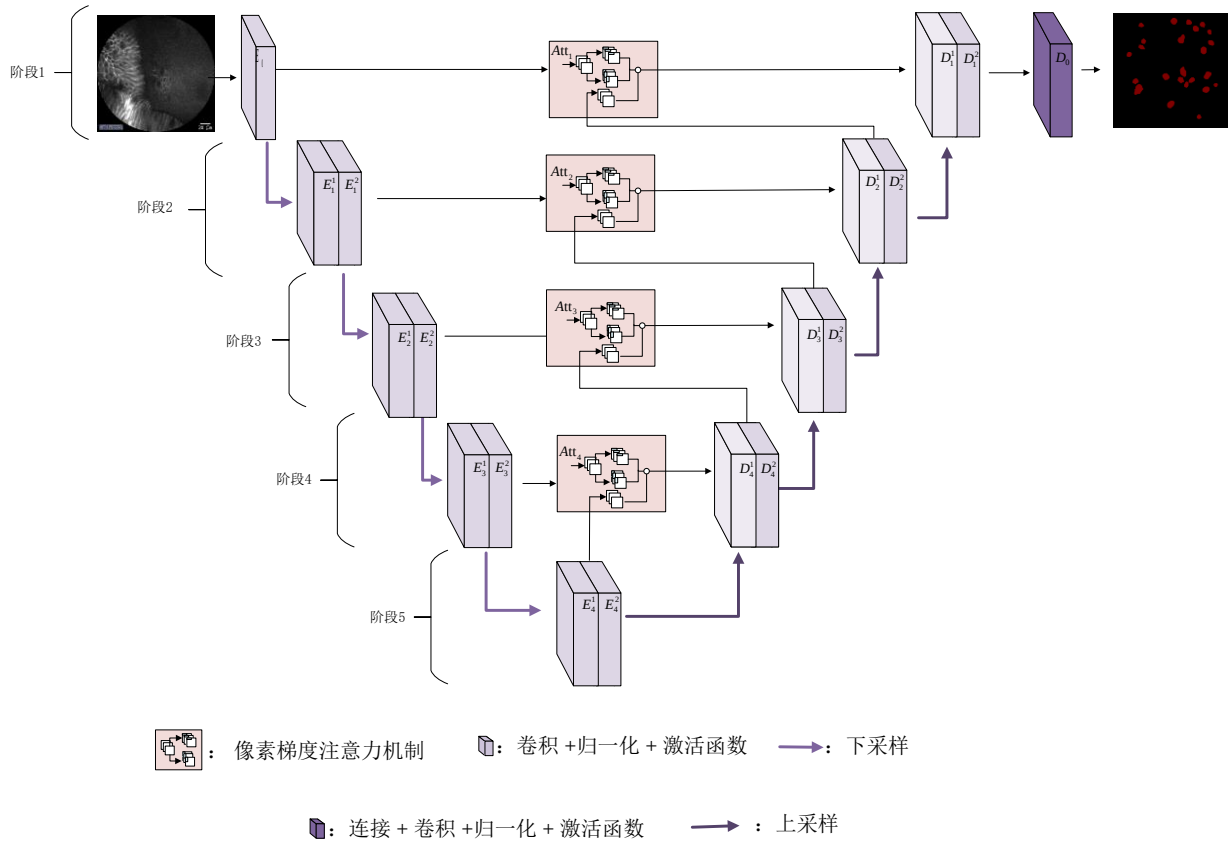

图 2-2 网络模型图

所提出的网络模型如图 2-2 所示，网络主要由编码层，同层注意力引导层以及解码层组成。网络的输入是$512 \times 512 \times 3$的三通道图像信息。在编码阶段，模型主要利用卷积操作对图像进行特征提取，并通过下采样来减小特征图的大小，以获得图像的深层特征。解码阶段则是利用像素梯度提取特征图与原始上采样特征图拼接后得到的深层网络特征进行杯状细胞区域识别。同层注意力引导层则是通过在浅层提取像素梯度特征信息，引

导模型上采样。最后，模型通过一个卷积层与分类函数将图像还原为$512 \times 512 \times 2$尺寸的概率图。

编码层是由激活函数与卷积核为 3×3 的卷积构成的。每个下采样模块包含两个 3×3 卷积核的卷积与一个激活函数层。如图 2-2 所示，原始数据在经过一个卷积核大小为 3×3，输入通道为3，输出通道为32的卷积后，进入下采样模块。每个下采样模块包含两个卷积核大小为3×3的卷积，与一个激活函数。GCSCLE选取relu作为激活函数，以便于梯度反向传播。当一个下采样模块完成特征提取后，通过线性下采样将特征图缩小为原来的二分之一，并将缩小后的特征传入下一下采样模块。通过此种级联的下采样操作，提取图像深层特征，降低模型训练压力。

在解码阶段，每一模块的输入为前一层上采样得到的特征图和像素梯度特征提取模块输出的特征图。通过两个大小为 3×3 卷积核的卷积与激活函数，提取图像特征。提取后的特征作为最终输出，传入下一阶段。最后，解码层将得到的特征向量输入一个输入通道为32，输出通道为2，卷积核大小为3×3的卷积中，并得到与原始图像尺寸相同的特征向量。特征向量通过 sigmod 函数进行分类后，得到最终的杯状细胞区域概率图，以确定杯状细胞区域。

2.3.2 像素梯度特征提取模块

目前，网络对目标区域大小的感知主要依赖于级联的卷积层。这种方法在训练上存在着一定劣势。级联卷积的劣势在于，当图像像素较小时，所需关注区域较小，模型较为容易训练。但是，当图像像素较大时，所需关注区域较大，级联的卷积结构在加深网络层数时需要大量的样本进行参数的训练。在胃黏膜肠化生阶段，胃黏膜肠上皮出现杯状细胞，在共聚焦激光显微内镜中呈现出大而黑的区域。然而，纤毛、染色质泄露等区域也会呈现出相同的颜色特性。相较于杯状细胞，此类区域在形状大小上有这明显区别。此外，受到染色影响，部分正常区域内会形成较小的灰黑色区域，这些区域也大大影响着网络的识别准确率。为此，GCSCLE 改进了门控注意力结构[58]，并将其引入同层连接层，用来提取识别图像的颜色梯度特征。

图 2-3 展示了像素梯度特征提取模块的详细结构。输入是来自编码层的参数E_i和来解码层的参数D_i。参数E_i来自相对较浅的特征，包含更多的像素信息，例如颜色、纹理、色

彩梯度等。因此，像素梯度特征提取模块分别通过在高度轴和宽度轴上的线性特征提取，来执行像素梯度特征提取。随后，通过归一化和激活函数，平滑获得的特征图。最后，该结构通过加性注意力机制进行高度轴和宽度轴上的特征融合。上述步骤如公式 2-1 所示。

$$A_{po} = Att_e(\mathrm{E}_i) = (\varphi_w(\sigma_w(W_w\mathrm{E}_i + b_w)) + \varphi_h(\sigma_h(W_h\mathrm{E}_i + b_h))) \quad (2\text{-}1)$$

其中，W_h是宽度轴上的特征提取，具体操作为采用1×7的卷积操作。σ_w和σ_h是对特征向量进行归一化操作。φ_w和φ_h是用于提取非线性特征的激活函数。b_h和b_w代表偏移向量。W_w是高度轴上的线性颜色特征提取，具体使用7×1的卷积操作。 A_{po}是输出的特征图。

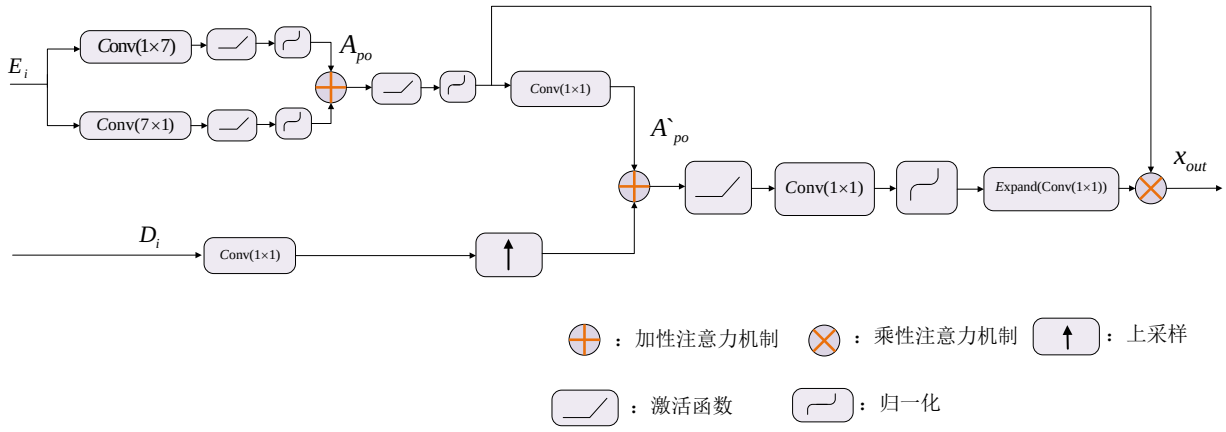

图 2-3 像素梯度特征提取模块

为了增强对目标信息的引导，GCSCLE 使用加性注意机制，使下采样的输入特征图D_i引导特征图A_{po}。GCSCLE 首先对输出特征向量A_{po}进行平滑操作，通过激活函数和归一化来获得向量$\widetilde{A_{po}}$。然后，像素梯度注意力机制对$\widetilde{A_{po}}$进行线性特征提取，并将向量通道放大到原始大小的两倍。随后，通过加性注意机制，将处理后的特征图与处理后的解码输入融合得到特征图。最后，对得到的特征图进行线性特征增强和上采样。上述步骤如公式 2-2 所示：

$$A_{po}{}' = Att_{po}(A_{po}, \mathrm{D}_i) = \left(\varphi_{po}\left(\sigma_{po}\left(W_pA_{po} + b_{po}\right)\right) + U(W_d\mathrm{D}_i)\right) \quad (2\text{-}2)$$

其中，W_d是目标的线性特征提取，用于特征增强。具体采用卷积核大小为1×1的卷积操作，输出通道数为输入通道的两倍。U是用于恢复特征图的线性上采样。W_p用于增强特征图。φ_{po}，σ_{po}和b_{po}分别表示层归一化、激活函数和偏移。

乘性注意机制通常比加法注意机制具有更高的灵敏度。为了进一步增强颜色梯度对

模型结果的影响，GCSCLE 将A_{po}经过激活和归一化处理，并通过乘法注意机制引导特征图$A_{po}{}'$。如图 2-3 所示，模型首先对特征图$A_{po}{}'$进行增强。随后，经过重采样与处理后的A_{po}融合。重采样操作采用线性方法将特征通道的数量减少到 1，然后扩展到原先大小。重采样操作有助于特征向量提取更多的低维信息。综上所述，像素梯度注意力机制如公式 2-3、公式 2-4 所示：

$$A_s = \tau(\varphi_s(w_s\left(\sigma_s\left(A_{po}{}' + b_g\right)\right))) \tag{2-3}$$

$$A_{out} = \delta\left(\varphi_{po}\left(\sigma_{po}\left(W_p A_{po} + b_{po}\right)\right), A_s\right) \tag{2-4}$$

其中，A_s是引导特征图，b_g和b_{po}是偏移量。σ_s是用于获得非线性特征的激活函数。w_s是用于提取线性特征的1×1卷积函数。φ_s和φ_{po}是特征激活函数。τ是恢复通道卷积运算。对于具体运算，通过卷积核大小为1×1的卷积将输入特征通道的数量减少到 1，然后通过叠加的方式将通道数量扩大至与与A_s通道数量相同。A_{out}是像素梯度注意力机制的最终输出。δ代表乘性注意机制。

2.3.3 损失函数

模型训练采用的损失函数为多分类的 Dice 损失。Dice 损失在医学图像分割任务中被广泛应用，能有效应对类别不平衡的问题。在我们的任务中，前景（杯状细胞区域）和背景（非杯状细胞区域）被赋予了对等的重要性。这样设计的目的是为了确保在训练过程中，网络能够同时关注杯状细胞和非杯状细胞区域，从而避免由于类别不平衡而导致的模型偏向某一类别的问题。

这种训练策略在处理较小的数据集时，能够放大预测结果与标注信息之间的差距。这种差距有助于网络更清晰地学习到不同类别之间的边界和特征，从而促进网络的快速收敛。在实际应用中，这样的损失函数设计可以显著提升模型的性能，使其在较小的数据集上依然能够取得良好的分割效果。损失函数定义为：

$$loss = 1 - (\frac{1}{2}\cdot\frac{2|X^0\cap Y^0|}{|X^0|+|Y^0|} + \frac{1}{2}\cdot\frac{2|X^1\cap Y^1|}{|X^1|+|Y^1|}) \tag{2-5}$$

其中，X^0为预测结果中的杯状细胞区域，Y^0为标注杯状细胞区域，X^1为预测结果中的背景区域，Y^1为标注结果中的背景区域。

2.3.4 算法描述

算法 2-1: 基于改进 U-net 的杯状细胞语义分割方法

输入: 杯状细胞临床图像$f_i = [3 \times \mathrm{H} \times W]$

//其中，3 为颜色通道。H 为宽 W 为高

输出: 杯状细胞区域$O = [\mathrm{H} \times W]$

//最终输出为杯状细胞区域 0-1 图，其中，1 为对应的杯状细胞区域，0 为背景区域。

输入图像重构 $f_r = reshape(3 \times 255 \times 255)$

// 重构图像将图像变成$3 \times 255 \times 255$的特征向量

$D_1 = \big(Conv(f_r)\big)$ //模型通过一次特征提取输入到改进的 U-net 中

for state **in** U_net_Down

// U-net 特征提取的编码阶段

$D_i = (DownSample(Conv\big(Conv(D_{i-1})\big)$

//编码过程中的特征提取

end for

for state **in U-net-up**

//U-net 特征提取的解码阶段

通过像素挤压得到特征向量A_{po}: A_{po} ←公式（2-2）

通过加性注意力机制得到$A_{po}{}'$：$A_{po}{}'$ ←公式（2-3）

经过特征提取后，得到引导向量A_{out} ←公式（2-4）

$E_i = Conv\big(Conv(Concat(A_{out}, Upsample(E_{i-1})))\big)$ //解码过程特征提取

end for

计算最终输出结果O:O ← $FCN(E_1)$

return O //返回输出结果

2.4 实验和讨论

2.4.1 评价指标

采用 Intersection over Union（IOU）、筛子系数（Dice）、精确率（Pre）、召回率（recall）、准确率（acc）以及 F2-score（F2）指标来综合评价分割结果。其中，交并比（IOU）定义如下：

$$IOU = \frac{X \cap Y}{X \cup Y} \quad (2\text{-}6)$$

其中，X为分割结果，Y为图像真实取值，$\cap$为X与Y的交点，$\cup$为X与Y的并集，IOU的取值为 0-1 之间，当 IOU 取值增大时，预测结果越接近真实结果。当IOU取值为 1 时，预测结果与目标结果完全重合。由此可以看出，IOU越大，分割效果越好。

筛子系数定义如下：

$$Dice = 2\frac{|X \cap Y|}{|X| \cup |Y|} \quad (2\text{-}7)$$

精确率定义如下：

$$Precision = \frac{TP}{TP+FP} \quad (2\text{-}8)$$

其中，TP 为模型预测为正样本中预测准确的像素数量，FP 为模型为负样本中预测错误的像素数量。

召回率定义如下：

$$Recall = \frac{TP}{TP+FN} \quad (2\text{-}9)$$

其中，FN 为模型预测为负样本的像素数量。

Accuracy 定义如下：

$$Accuracy = \frac{TP+TN}{TP+TN+FN+FP} \quad (2\text{-}10)$$

F2 指标定义如下：

$$F2 = \frac{5Precision \cdot Recall}{4Precision+Recal} \quad (2\text{-}11)$$

2.4.2 实验细节

实验在 Linux 系统上进行，通过开源框架 Pytorch[60]进行搭建。实验在显卡 NVIDIA GTX 3090i GPU （with a 24-GB memory）与一个 12-core PC with an Intel （R） Core （TM） i9-7900X CPU @ 3.30 GHz 3.31 GHz （with 16-GBRAM）处理器上完成。

实验通过8: 2比例划分数据集数据，将 334 张图像中的 267 张作为训练数据，67 张作为测数据。通过图像放缩，将所有图像尺寸设置为512×512。批处理次数设置为 4，最

大 epoch 为 200，学习率（lr）设置为 0.001。选取训练集上准确率最高的结果作为参数初始化模型。

2.4.3 实验结果

本节将 U-net[25]、U-net++[32]、Att-unet[71]、seg-Net[28]与 GCSCLE 模型进行对比。表 2-1 中列出了上述算法的 IOU，Dice，Precision，Recall，Accuracy，F2 等评价指标。如表 2-1 所示，本节首先实现了 U-net 在杯状细胞数据集上的语义分割。U-net 通过同层连接将浅层特征传递至与深层，并通过上采样与卷积操作进行语义分割。表中反映出，U-net 的 IOU 达到了 83.35%，精确率仅为 81.36%，远低于其他网络。相较于 U-net 模型，U-net++通过引入密集连接在交并比以及召回率上有一定提升，达到了 84.00%以及 86.13%。但精确率却有较大下降，仅为 78.02%。这是由于少量的数据在训练大规模参数时，无法关注到关键特征。这就使得模型容易将杯状细胞区域与部分正常细胞区域混淆。

表 2-1 实验数据结果

	IOU	Dice	Precision	Recall	Accuracy	F2
U-net[25]	0.8345	0.8050	0.8136	0.8122	0.9895	0.8073
U-net++[32]	0.8400	0.8123	0.7802	0.8613	0.9897	0.8393
seg-Net[28]	0.8402	0.8127	0.8227	0.8167	0.9901	0.8136
Att-unet[71]	0.8552	0.8345	0.8350	0.8399	0.9907	0.8370
GCSCLE	0.8795	0.8664	0.8554	0.8834	0.9925	0.8758

为对比模型识别效率，本节还实现了 seg-Net 网络模型。seg-Net 网络模型通过计算编码器中的池化索引，进而优化上采样过程，并提供更精确的结构重建。实验现实，seg-Net 的改进在一定程度上能优化识别结果，精确率达到了 82.27%。同时，引入注意力机制的 Att-unet 在各项指标上均好于其他算法，IOU 达到了 85.52%，精确率达到了 83.50%。但是，Att-unet 所提出的同层特征引导仅仅经过1×1卷积操作，这种方式对识别结果的提升仍有巨大改进空间。

从表中可以看出，GCSCLE 在 IOU 上达到了 87.95%，远高于其他网络模型，且在 Dice，Precision、Recall 上均有一定提升。这是由于，GCSCLE 在同层连接过程中加入像素梯度特征提取模块，网络在关注到颜色梯度特征的同时，也能关注到边界、形状等特

征。这使得 GCSCLE 在兼顾目标区域大小的同时，关注到了像素梯度变化，进而大大提升了区域识别的准确性。第 2.4.4 节将对此展开具体分析。

2.4.4 讨论

2.4.4.1 分割结果分析

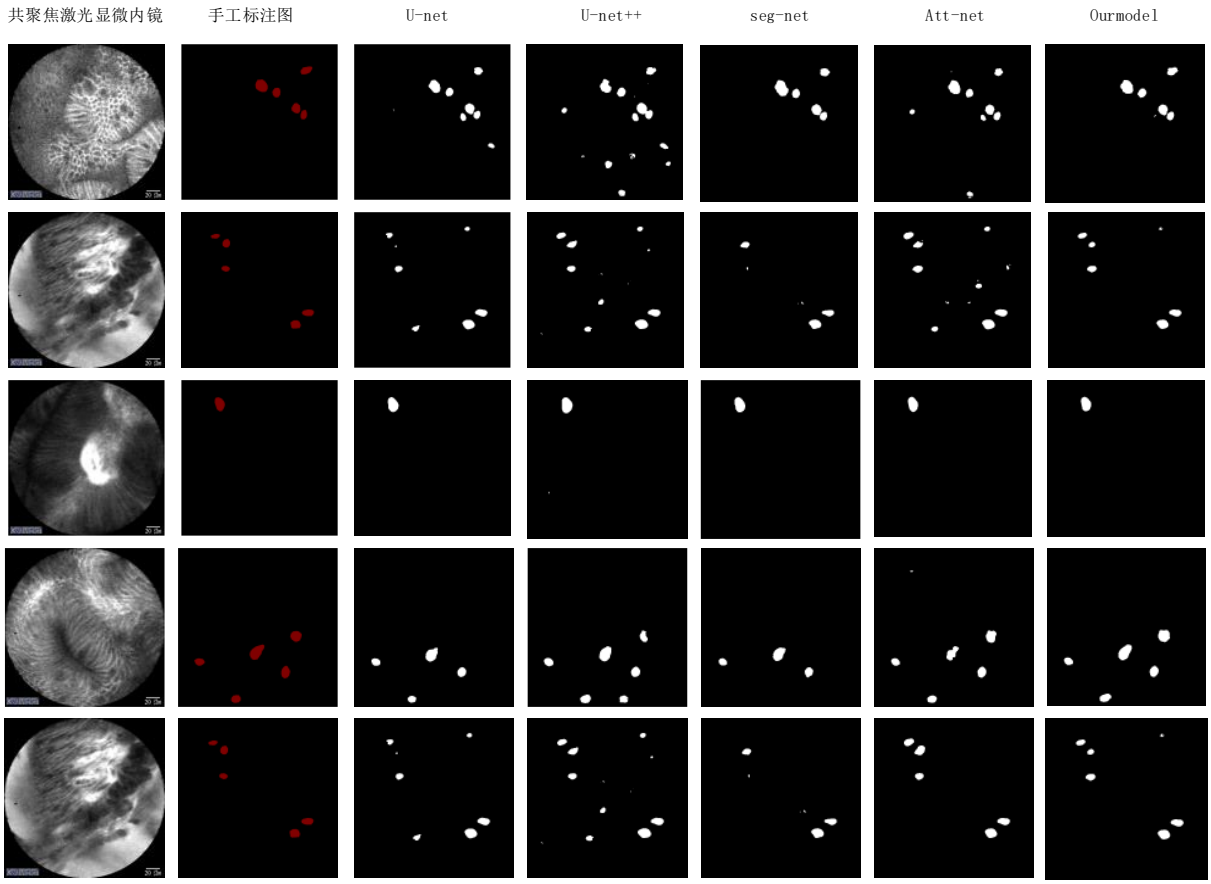

图 2-4 模型效果对比图

图 2-4 展示了不同模型在共聚焦激光显微内镜图像上的杯状细胞语义分割结果。如图 2-4 所示，U-net 模型与 U-net++模型比较容易将正常区域误识别为杯状细胞区域。seg-Net 网络则无法关注到杯状细胞区域的边界信息。Att-net 虽然在一定程度上优于其他算法。但当图像较为复杂时，仍无法较好的排除正常区域。特别的，GCSCLE 在共聚焦激光显微内镜数据集上展现了巨大的优势。如图所示，在分辨率较大的共聚焦激光显微内镜图像上，模型在准确提取颜色特征的同时能有效关注像素梯度信息。当面对细胞质泄露、纤毛区域等复杂情况时，模型仍能展现出可靠的分割结果。

2.4.4.2 模型分割结果与人工标注结果对比

图 2-5 对比了模型识别区域与人工标注区域。如图 2-5 所示，红色区域为人工标注但

未被该方法识别的区域。白色区域为 GCSCLE 识别但人工未标注的区域，青色区域则是两者的重叠区域。如图，GCSCLE 识别的杯状细胞区域接近人工识别结果。然而，人工标注需要大量人员的反复校对，费时费力。GCSCLE 在速度上远优于人工识别，且接近真实人工标注信息。

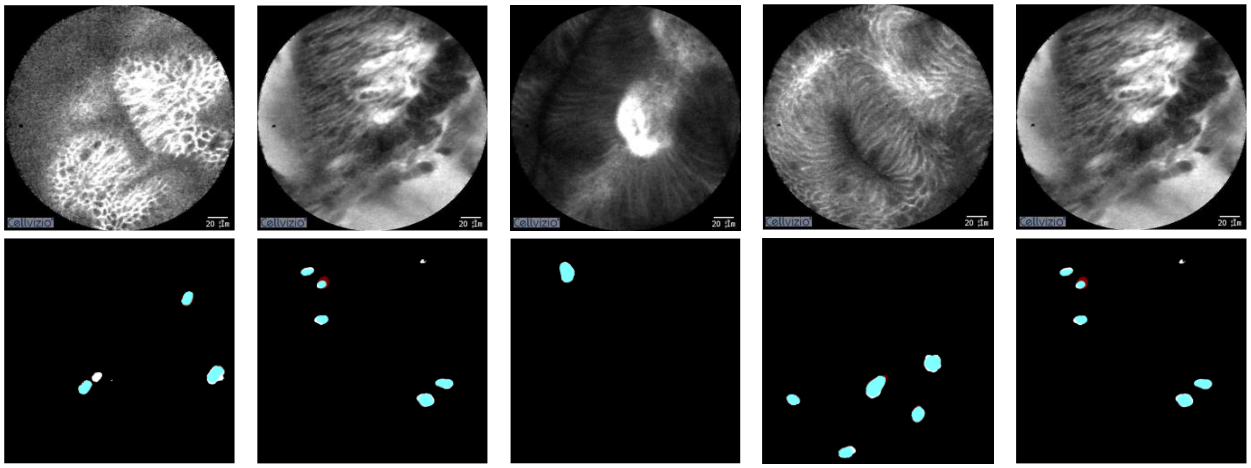

图 2-5 共聚焦激光显微内镜与人工标注区域对比。其中，白色为模型识别区域但非标注杯状细胞区域，青色为模型识别与人工标注重叠区域，红色为模型识别区域但人工非标注区域

2.4.3.3 箱图分析

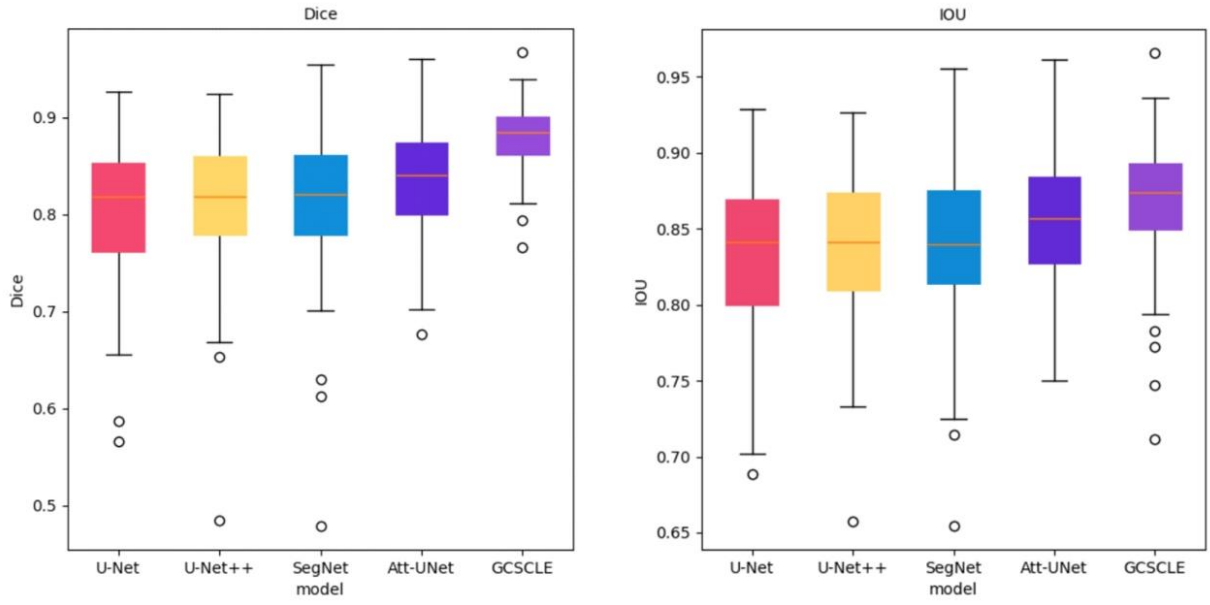

图 2-6 不同模型在验证据上的 Dice 与 IOU 箱图

图 2-6 展示了不同模型在杯状细胞语义分割任务上的 Dice 与 IOU 箱图。箱图从左到右依次展示了 U-net[25]、U-net++[32]、seg-Net[28]、Att-unet[71]、GCSCLE 的分割结果分布。图中橙线代表模型识别结果的中位数，圆圈表示异常值。如图 2-6 所示，与其他模型相比，U-net 的分割结果分布较为分散且识别准确率较低。其中，Dice 和 IOU 主要分别在 0.76-

0.87之间。与U-net相比，U-net++和seg-Net在Dice系数和IOU方面均有一定的改进，相较于U-net提升2-3个百分点。图中反映出，U-net++识别结果较为集中，具有较高的鲁棒性。seg-Net分割结果的IOU跟DICE则较为分散。图中同样反映出，Att-unet在IOU和Dice的分割结果上有明显改善，但识别结果较为分散。这反映出Att-unet无法应对一些复杂情况。

从图2-6中可以看出，GCSCLE在杯状细胞语义分割任务上表现更为稳定，识别出的结果分布更为密集。这反映出GCSCLE在应对临床不同状况时，能更为有效的捕获杯状细胞特征，能够在复杂的情况下更为精准的识别杯状细胞区域。

2.5 本章小结

本章提出了基于改进U-net的共聚焦激光显微内镜杯状细胞语义分割方法。该方法利用基于改进U-net的共聚焦激光显微内镜杯状细胞语义分割模型，获取共聚焦激光显微内镜中杯状细胞区域。相较于现有的模型，改进的U-net将像素梯度特征提取模块融入同层链接过程中，并将得到的特征图引导模型上采样过程。通过此种特征引导方式，模型能进一步提取像素梯度特征，区分杯状细胞与正常细胞区域，提升分割准确率。为了验证GCSCLE的优越性以及实用价值，本章在共聚焦激光显微内镜杯状细胞语义分割数据集上进行了分割结果分析，标注结果对比等一系列实验。实验证明，GCSCLE在保证较高的准确率同时，分割结果与人工分割相差无几。综上所述，这项工作为结合深度学习的肠化生诊断提供了很好的启发，也为胃肠道病变的诊断提供了可能的研究方向。未来，这项研究在共聚焦激光显微内镜检测取代活检、共聚焦激光显微内镜肠化生严重程度评估等研究中有着重要意义。

第三章 结合内窥镜视频的共聚焦激光显微内镜检测部位识别方法

3.1 引言

Correa序列指出[61]，胃癌是由慢性胃炎、萎缩性胃炎、肠上皮化生、上皮内瘤变发展而来的。其中，胃黏膜肠上皮化生是肠型胃癌发生的危险因素之一，被视为是重要的癌前病变。临床上，共聚焦激光显微内镜是检测肠化生严重程度的关键步骤。Guo等人[62]的研究证明，通过共聚焦激光显微内镜提取胃窦、胃体和胃角处的细胞结构信息，能有效识别患者肠化生病变程度，避免活检带来的二次损伤。同时，共聚焦激光显微内镜的成像原理与活检采样原理类似，具有较高的医学参考价值。目前，共聚焦激光显微内镜检测已被应用于多种胃部疾病的诊断（如肠化生[63]、胃癌[64]、巴雷特食道癌[65]等）。

临床常用的共聚焦激光显微内镜[66-67]为探头式共聚焦激光显微内镜。该设备将共聚焦激光镜头单独做成探头，直径小、清晰度高，可以从内窥镜活检通道插入，与多种型号内窥镜均可搭配使用。具有扫描速度快、成像质量高等优势。如图 3-1 所示，探头式共聚焦激光显微的检测过程是共聚焦探头通过内窥镜活检通道进入患者体内。影像科医师通过观察内窥镜成像，控制探头在患者体内的移动，捕获患者体内的微观信息。在实际操作中，当共聚焦探头贴近胃壁时，呈现出清晰的微观结构成像，反映患者病情特征。当扫描完成时，医生抬起内镜探头，该位点扫描结束。从贴近到抬起的这段视频序列，则是反映患者病情的信息帧（Informative Frame，IF）序列。

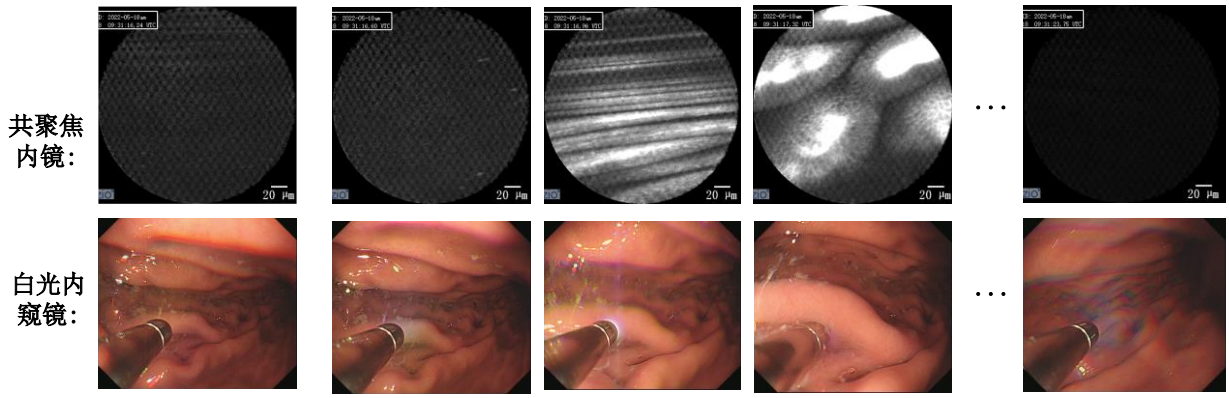

图 3-1 相同时刻内共聚焦激光显微内镜以及胃内窥镜对应关系图。

相较于常规的内窥镜等宏观检测方法，共聚焦激光显微内镜的微观成像结构无法体现检测部位信息，医生无法实时判断检测部位。这会造成治疗分析过程中发病部位难以

确定、无法实现靶向治疗等问题。另一方面，在后续诊断过程中，内窥镜视频和共聚焦激光显微内镜视频往往存储在不同设备。医生不得不综合分析内窥镜视频以及共聚焦激光显微内镜视频片段（如图 3-1 所示）来分析发病部位以及严重程度，造成了观察上的困难。为此，本文研发了共聚焦激光显微内镜检测部位识别方法。通过实时识别检测部位，能有效降低漏诊率，保证检查覆盖率和检查质量，减轻医生工作负担。

在共聚焦激光显微内镜检测过程中，人工操作的探头移动很难做到匀速控制。临床检测过程中的部位跳跃、信息变换等情况会导致模型在相同视觉语义线索内赋予帧不同的涵义。例如，在共聚焦激光显微内镜探头处于胃窦与食管交界处时，共聚焦探头向上移动则为探索胃窦大弯区域，向下则为胃窦小弯区域。现有神经网络很难做到该类信息的精准感知。目前，关于视频注意力的研究通常侧重于采用上下文的特定方面（通道、空间、时间或全局上下文等）来细化特征。忽略了在计算注意力时的潜在相关性，导致不完整的上下文利用，进而忽略了视频在空间上的运动聚合关系。换而言之，单帧下注意图的生成是一次性的表达，没有考虑到不同语境之间的相关性。这种处理方式在面对气泡、胃部蠕动、粘液等情况时，模糊帧（受到气泡、胃部蠕动、粘液影响较大的帧）会大大影响识别准确率。造成模型识别误差的产生甚至识别上的错误。另一方面，胃部空间较小，相较于自然运动图像，共聚焦激光显微内镜探头移动速率分布严重不均。跨部位的长距离位移、局部小范围移动交错分布于检测过程中。这就导致了时序特征难以提取、空间信息难以掌握。

为降低部位漏诊概率、保证检查覆盖率，本研究根据相关解剖学构造，将胃部部位分成了胃窦大弯、胃窦小弯、胃体大弯、胃体小弯以及胃角。提出了一套用于共聚焦激光显微内镜检测部位识别的方法（Probe-based Confocal Laser Endomicroscopy Diagnosis Area Identification Method，pCLEDAM）。方法通过融合检测过程中的内窥镜成像以及共聚焦激光显微内镜成像，实时识别共聚焦激光显微内镜检测部位。具体而言，pCLEDAM 通过检测共聚焦激光显微内镜的工作状态，提取信息帧序列前 1.6 秒内的内窥镜成像视频。视频通过深度网络模型，识别共聚焦激光显微内镜的检测部位。此外，针对现有网络对于关键特征难以提取、时序信息关注度不足的问题，本章提出了共聚焦激光显微内镜检测部位识别模型。该模型通过融合相邻帧提取单帧关键特征，通过激励—挤压模块促进网络模型关注时序变化。实验显示，pCLEDAM 能准确识别共聚焦激光显微内镜检测部位，有

效保障检测覆盖率。

3.2 材料

受到移动过程中的视野范围、胃部蠕动干扰等因素限制，临床仅使用某一帧来识别共聚焦激光显微内镜的检测部位是不可靠的。为此，本研究从齐鲁医院收集了 67 例共聚焦激光显微内镜临床诊断案例，经过相关专家筛选以确保规范性。每个数据样本持续时间约为 10-20 分钟，包含一段共聚焦激光显微内镜成像视频以及内窥镜视频信息。这些临床诊断案例被分为一个训练集（47 个临床案例）和一个测试集（15 个临床案例）。根据共聚焦激光显微内镜检测信息，采样前的 1.6 秒内的内窥镜视频并进行存储。经过人工核验后，共采集有效实验样本 500 段。内窥镜视频均在 OLYMPUS EVIS LUCERA ELITE CLV-290SL 的白光内窥镜下拍摄，FPS 为 25，分辨率 1920 × 1080 /帧。删除对个人信息（如检查日期、患者姓名等），以确保隐私和安全。在相关分类标准的指导下，三位专家被邀请对采集到的内窥镜视频进行标注。标记信息划分为 5 个不同的解剖部位，反复校对以确保实验数据的可靠性。

3.3 方法

3.3.1 方法概述

本文提出的结合内窥镜视频的共聚焦激光显微内镜检测部位识别方法如图 3-2 所示，方法利用了共聚焦激光显微内镜工作状态以及内窥镜视频信息，实现了共聚焦激光显微内镜检测部位识别。如图所示，帧状态监听作用在共聚焦激光显微内镜上，实时检测共聚焦激光显微内镜的工作状态。工作状态的检测通过 Resnet 网络[72]、色阶分析等方案便可得到准确识别，在此不做过多赘述。具体而言，系统实时监听探头成像的工作状态。若检测到由非工作状态转换为工作状态时，判断距离上一段信息帧序列间隔时间是否大于 1.6 秒。若满足条件，将状态超参数 status 传递到内窥镜中。系统开始调用内窥镜视频分析模块。否则，系统继续监听共聚焦激光显微内镜的工作状态。

内窥镜视频分析模块的作用是基于内窥镜视频信息，识别共聚焦激光显微内镜的检测部位。如图 3-2 所示，当内窥镜收到模型检测状态码时，回溯 1.6 秒内的内窥镜视频信

息。回溯到的内窥镜视频通过固定间隔抽帧的方式抽取 10 帧图像输入共聚焦激光显微内镜检测部位识别模型中（模型参数由调参、训练所得）。模型根据输入的帧序列识别共聚焦激光显微内镜的检测部位，并将识别结果传递给共聚焦激光显微内镜设备。共聚焦激光显微内镜更新当前部位参数，并继续监听共聚焦激光显微内镜检测影像。

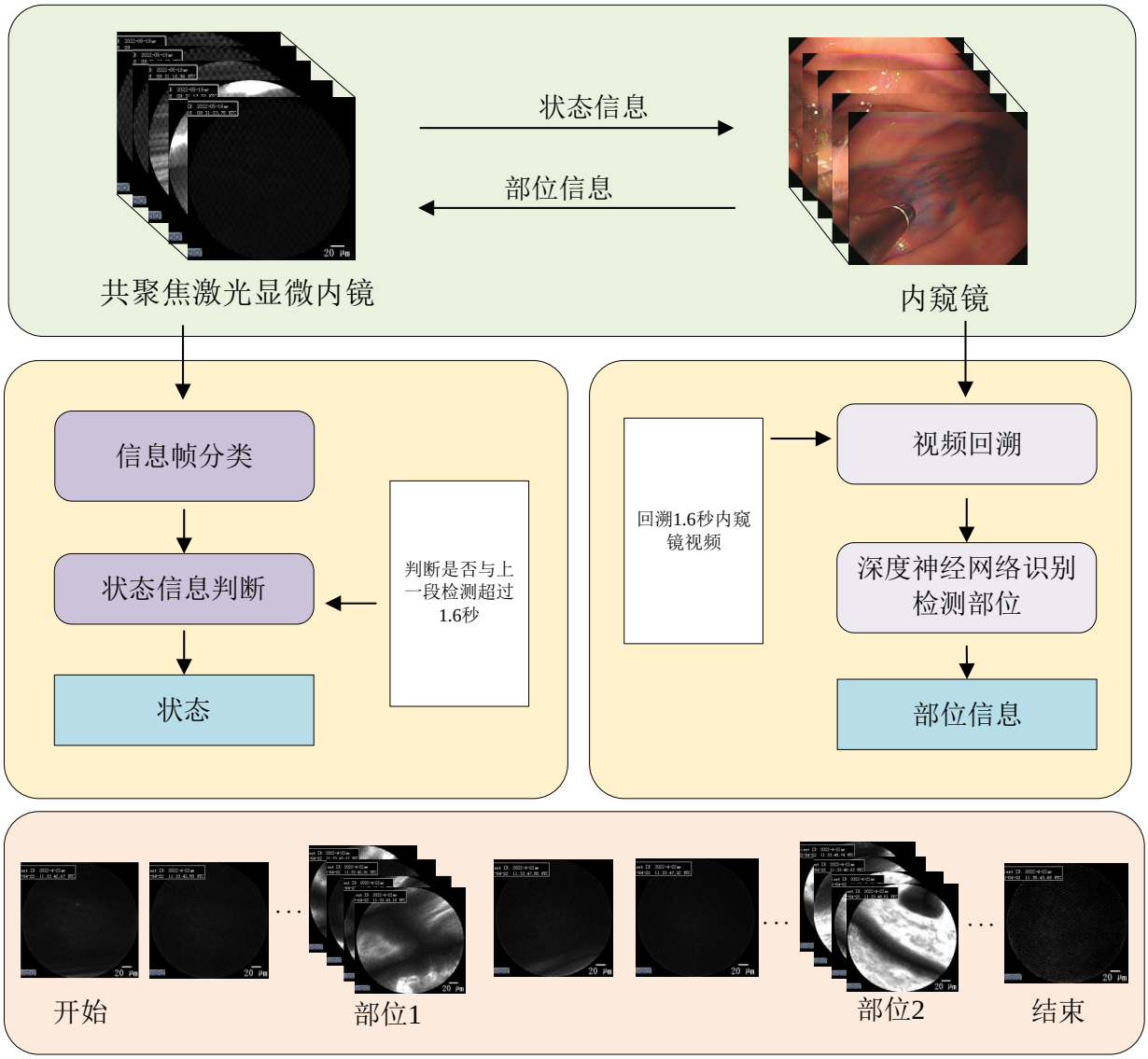

图 3-2 共聚焦激光显微内镜检测部位识别结构图

3.3.2 共聚焦激光显微内镜检测部位识别模型

如图 3-3 所示，提出的共聚焦激光显微内镜检测部位识别模型主要有两部分组成。输入的视频序列先后经过基于沙漏卷积的单帧关键特征捕捉模块和时序特征敏感的空间特征网络模块进行特征提取。对于提取到的特征，模型通过全连接神经网络进行分类计算，并得到最终的分类结果。模型组成包括：

（1） 模型的输入为共聚焦激光显微内镜起始时间前 1.6s 的内窥镜视频帧序列。通过固定时间间隔的帧抽取以及图像变换，得到特征向量f：$f = R^{10\times3\times224\times224}$，10 代表的抽取到的 10 张帧图像，3 为 RGB 特征通道，224 为压缩后图像的宽与高。

（2） 基于沙漏卷积的单帧关键特征捕捉由定长滑块以及沙漏卷积组成。对于输入的特征f，通过窗口大小为$5 \times 3 \times 224 \times 224$的定长滑块依次截取单帧及相邻帧的特征向量。截取到的特征向量通过双层沙漏卷积提取滑块中心帧的关键特征（例如胃窦口、探头位置等）。

（3） 随后，特征被输入到时序敏感的空间特征模块中，通过改进的Resnet提取时序特征。具体而言，模块将时序偏移以及空间注意力挤压机制引入每一个 Resnet 的 block，增强网络对时序特征的感知能力。输入的特征通过该模块后，得到最终的视频特征图。

（4） 最后，将得到的特征图展开后输入到FCNN中，预测视频检测部位。检测部位包含胃窦大弯、胃窦小弯、胃体大弯、胃体小弯以及胃角。

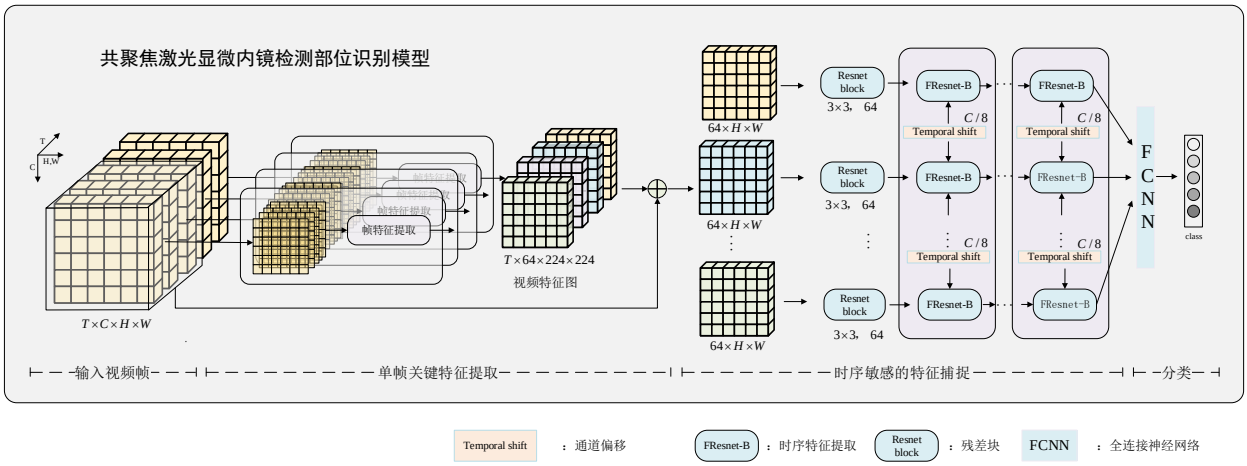

图 3-3 基于沙漏卷积和时序特征的共聚焦激光显微内镜检测部位识别模型

3.3.3.1 基于沙漏卷积的单帧关键特征捕捉

在共聚焦激光显微内镜的检测过程中，共聚焦激光显微内镜探头跟随着内镜探头在胃部三维空间内移动。移动过程中，贲门口、共聚焦探头等关键信息在成像上存在着大小以及形状上的变换。Tan 等人[68]证明，基于刚性的卷积针对此类信息很难做到有效的特

征捕捉。为此，pCLEDAM 将沙漏卷积[68]运用到单帧关键特征的提取过程中。通过不同接受场的特征提取，促进模型关注到关键信息的形状、位置变化（如图 3-4（b）所示）。沙漏卷积如公式 3-1 所示，对于特定的时间 K，沙漏卷积的特征提取是通过不同大小的接受场。通过偏移量以及不同的大小的卷积核，实现接受场大小的改变。

$$HgC(X)_{t,h,w} = \sum_{i=-\frac{K}{2}}^{\frac{K}{2}} a, f(X_{t+1}; W_{p|i|+1,p|i|+1})_{h,w} \tag{3-1}$$

其中，f表示空间聚合操作，$W_{p|i|+1,p|i|+1}$表示扩展接收字段的大小。

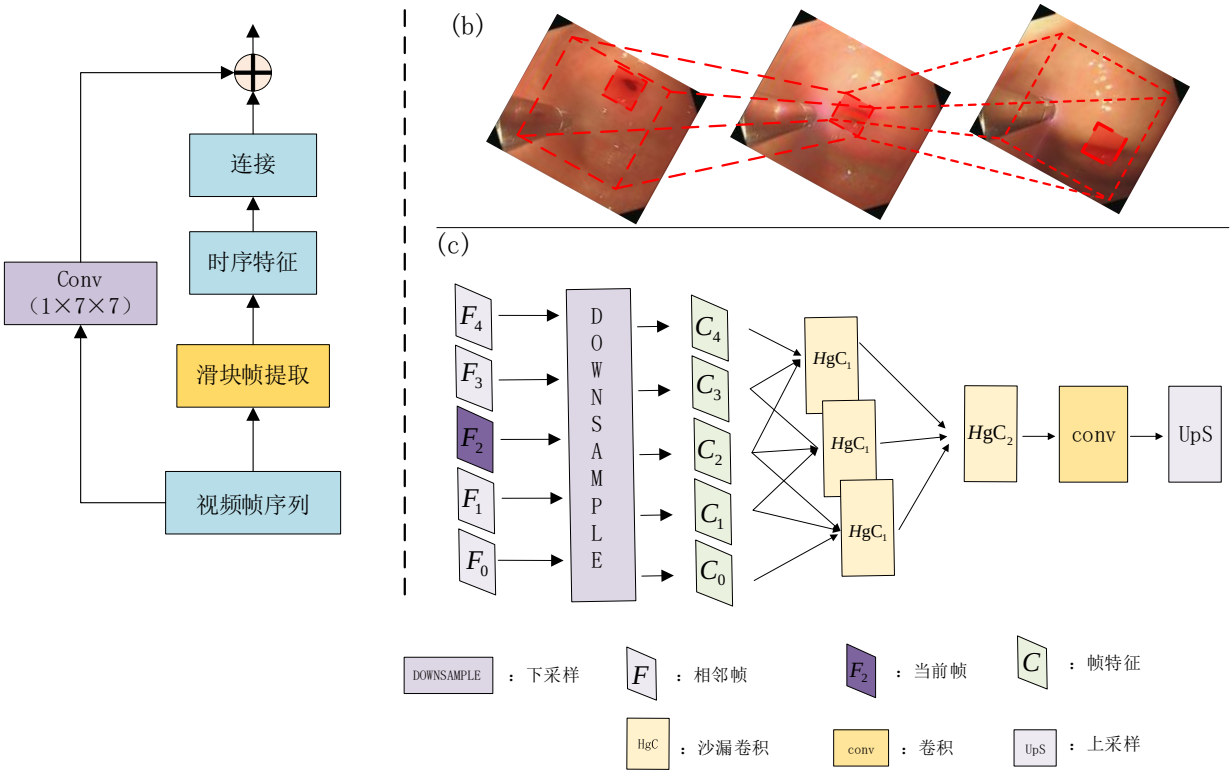

图 3-4 单帧动作特征建模

引入沙漏卷积的单帧关键特征捕捉如图 3-4（a）所示，由一个定长滑块以及基于双层沙漏卷积的特征提取组成。首先，对于输入的特征$F \times 3 \times H \times G$，基于沙漏卷积的单帧关键特征通过定长为 5 的滑块，逐步截取当前帧及其前后相邻 2 帧（共 5 帧）的特征向量。截取到的特征向量为$5 \times 3 \times H \times G$的特征图。其次，pCLEDAM 采用了 2 层沙漏卷积对截取到的张量进行单帧运动建模，以提取单帧关键特征。当滑块滑动到输入序列末端时，模型将提取到的特征连接得到包含单帧关键信息的特征序列。最后，特征序列通过加性注意力机制，与经过卷积核为$1 \times 7 \times 7$的 3D 卷积提取到的初始特征进行融合。特别的，对于起始帧特征，pCLEDAM 通过均值补差，增添了首两帧。这是为了在不破坏视频特征

的情况下，将输入特征与模型提取特征进行融合。结尾帧采用类似操作补齐。

单帧运动建模如图 3-4（c）所示，输入信息序列经过下采样后，依次输入沙漏卷积HgC_1以及HgC_2中，得到蕴含单帧关键信息的特征向量$H_t^{`}$。其中，下采样采用线性下采样以降低模型训练压力。$HgC_1$的输出向量维度为3，$HgC_2$输出向量维度为9。通过双层沙漏卷积，识别模型关注到关键信息在形状、大小上的改变。随后，$H_t^{`}$通过一个卷积以及上采样，将特征向量的H，W还原至输入状态，以实现与输入向量的融合。单帧运动建模如公式 3-3 所示：

$$H_t^{`} = HgC_2\left(HgC_1\left(DownSample(C_t)\right)\right) \tag{3-2}$$

$$H_t^{fm} = Upsample\left(Conv2d\left(\left(H_t^{`}; 7\times 7\right)\right)\right) \tag{3-3}$$

其中，$DownSample$函数为下采样操作，$UpSample$则为上采样操作。两个HgC，即$HgC1$和$HgC2$ 为沙漏卷积，用以实现关键特征捕捉。

3.3.2.2 时序特征敏感的空间特征提取

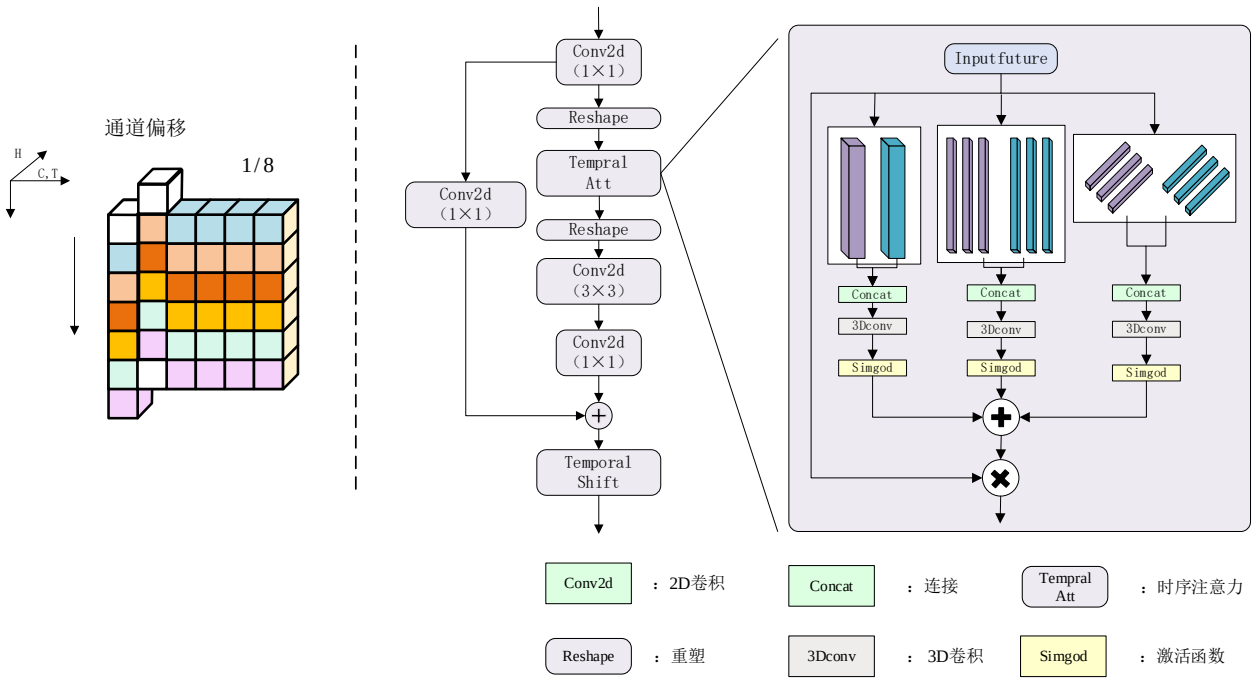

图 3-5 时序特征敏感的空间特征提取结构图

传统的 3D 卷积内存消耗巨大，无法实现临床实时预测的目标。此外，3DConv 难以提取时间序列上的关键特征，造成时序信息分类上的误差甚至错误。TSM 模型提出的时序特征提取采用基于通道偏移的注意力机制，并证明了该机制在时间模块上提取的有效