**技术交底书**

**发明名称： 一种基于多模态大模型提示学习机制的前列腺癌诊断方法**

**本专利发明人： 向莉华 倪张凯 朱俊泽 ­**

**技术交底书撰写人及技术联系人：\_倪张凯**

**电话： 13646019080 FAX: \_ \_ E-MAIL：\_zkni@tongji.edu.cn \_**

注意：

1.**第2部分和第4部分**，**一定要写的全面、清楚**。

2.英文缩写要有中文译文和英文原词，最好在术语解释部分给出。

3.**全文对同一事物的叫法应统一，避免出现一种东西多种叫法。**

4.应该阐述发明目的是通过什么技术方案来实现的，不能只有原理，也不能只做功能介绍。

**术语解释：**

解释一些跟技术方案有关的专业术语

1. **本发明要解决的技术问题是什么？**（务必明确一个最主要的技术问题）

前列腺癌症超声视频影像对比语言-图像预训练方法

**2、详细介绍技术背景,并描述已有的与本发明最相近似的实现方案**（与本发明最接近的技术方案的说明（对于方法，应说明现有方法的步骤，对于装置，应当说明结构组成及其关系））

任务解释

前列腺癌是全球男性中第二常见的癌症类型，也是导致癌症相关死亡的主要原因之一。在《中国前列腺癌筛查与早诊早治指南（2022年）》中，强调了提高筛查效果和规范性。尤其是在低资源地区，低成本且高效的诊断方法尤为关键。作为现代医学诊断的重要工具，医疗影像技术广泛应用于疾病检测中，并发挥着不可替代的作用。医疗影像技术包含磁共振成像（Magnetic Resonance Imaging, MRI）、计算机断层扫描（Computed Tomography, CT）、X光成像以及超声成像技术等。这些技术提供了宝贵的视觉数据，帮助医生诊断和评估疾病。在前列腺癌的诊断中，MRI是最常用的影像学工具。MRI具有较高的特异性，可用于前列腺癌的T分期预测，但其敏感性较低，且受制于价格、禁忌症及设备普及等因素，难以普及到广泛的临床应用中。作为另一种常见的影像诊断工具，超声成像在前列腺癌检测中也有所应用。与MRI相比，超声成像具有成本低、适用人群广和对人体危害小的优势，更适用于低成本且高效的前列腺癌检测。但前列腺癌多发、散灶的特征使得单张影像的诊断价值降低，包绕整个前列腺腺体及周围组织、器官的超声视频弥补了单张影像诊断的不足。然而超声视频成像存在分辨率低的问题，导致难以准确进行诊断。此外，超声的敏感性和特异性存在较大波动，尤其在小病灶和边界模糊的情况下，诊断的准确性依赖于医生的个人经验。因此，如何有效提升超声视频的分辨率，以提供更加清晰且高保真的影像，并提高医生对小病灶及边界模糊等复杂区域的诊断准确性，已成为当前基于超声成像开展前列腺癌诊断的重要研究方向和技术发展重点之一。

在前列腺癌的诊断中，超声视频作为一种重要的成像工具，虽然具备了低成本、适用人群广泛等优点，但由于其分辨率低和诊断准确性受限，特别是在处理小病灶或边界模糊的区域时，仍然面临许多挑战。因此，如何提升超声视频的诊断能力，特别是在处理复杂病灶区域时，已经成为了前列腺癌超声诊断领域的核心问题之一。视频分类任务在这一背景下成为了提升超声视频诊断效果的重要手段之一。与传统的单张静态影像不同，超声视频通过连续的帧信息提供了时序特征，这对于捕捉前列腺癌的动态变化、病灶发展及其与周围组织的关系等具有重要意义。通过对超声视频中的多个帧进行分析，可以更全面地评估前列腺组织的结构、血流情况以及肿瘤的形态等，从而提高疾病的检测率和诊断的准确性。

近似实验方案

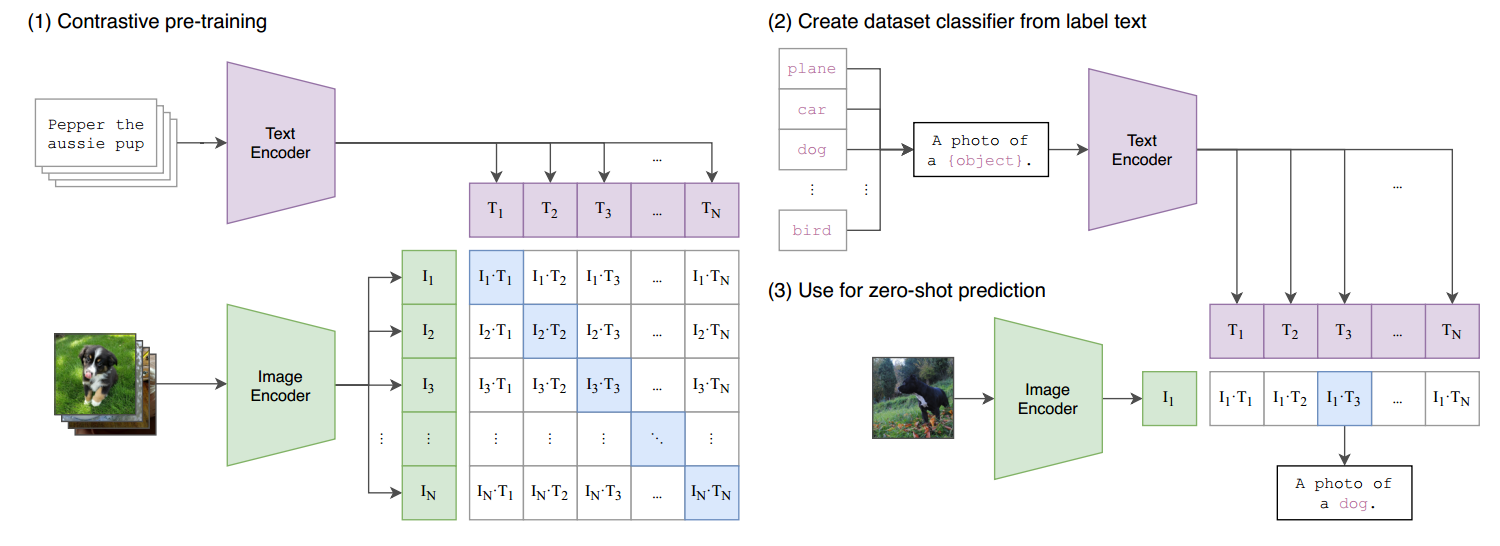
****

图2.CLIP预训练流程框架

基础图片-文本对齐（Contrastive Language-Image Pretraining,CLIP）的方法如图1所示。它为医学信息和医学图像对齐提供了一种对齐思想，应用到视频模态中需要提取每一帧和文本做对齐，对于视频特征使用残差网络（ResNet）或者视觉转换器（Vision Transformer），I代表输入的图像，VI代表提取出的特征，维度为。批次大小(Batch Size)，H为图片高度，W为图片宽度，C为特征通道数。

在文本特征的提取中，使用预训练的文本编码器，T代表图像对应标签的文本，VT表示获取的文本特征，特征维度为。n为批次大小(Batch Size)，L为文本长度。

获取两个模态的特征后通过对比学习拉近对应图像和文本的相似度，其中S(T,I)表示文本T和图像I的相似度，VI⋅VT代表向量点积，||VI||和||VT||为各自向量的欧几里得范数。最后通过InfoNCE损失更新参数。

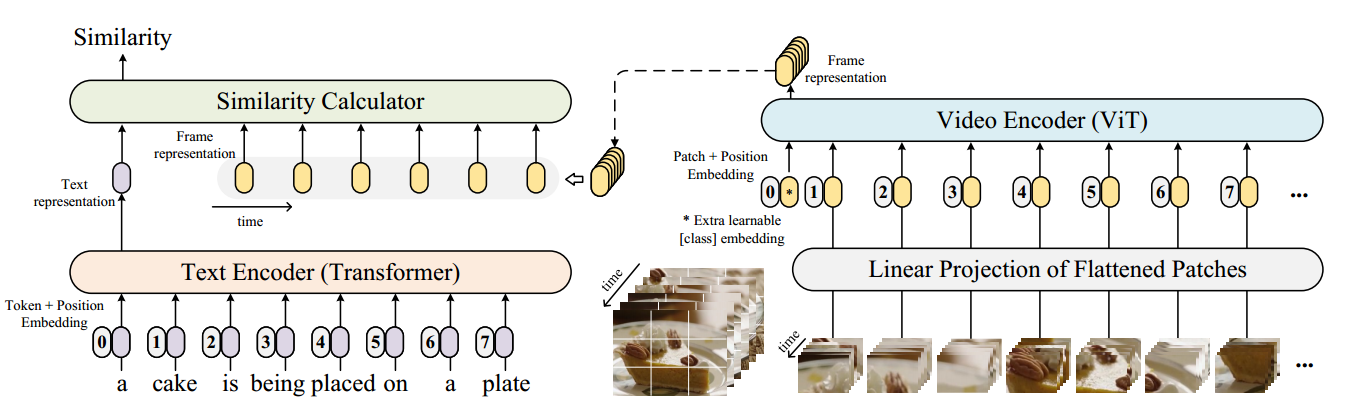


图2.视频CLIP预训练流程框架

那么在视频中的流程如图2参考，考虑对t帧采样，对一个视频video 其中T代表采样的帧数，H和W代表视频帧的尺寸，按照视觉转换器（Vision Transformer, ViT）的要求要把向量按照的大小扁平化成N个向量,加入了额外的分类令牌Xcls之后整个每帧即将输入ViT的令牌(Token)如下

其中e代表时间和空间的位置编码。Pemb代表线性投影层的参数。之后取出类别令牌（Class Token）作为图像表征并按照时序排序和文本特征做特征相似度对比。

**3、以因果关系推理的方式推导出现有技术的缺点是什么？**

**针对这些缺点，说明本发明的目的。**（缺点可以有多个，但一定要分清主次。最主要的缺点原则上只能有一个。）

其中针对前列腺癌症超声影像的数据方面有如下问题：

1. ：数据需要医生进行逐帧标注，劳动力和时间成本高昂。反之如果标签过粗，仅对整个视频执行描述，模型学习到的特征和文本描述对齐的效果就不是非常的理想。在前列腺的超声影像数据采集的过程中，有相对长的寻找病灶部位的过程，这个过程的无关信息被模型捕捉，这造成的一定的干扰。这是最主要的缺点。

为了解决上述问题，本发明提出了一种基于大语言模型丰富基础知识的可泛化的前列腺癌症T期诊断系统。该方法不需要医生额外的对视频中的每一帧图像进行标注，通过大语言模型分割能力，设计了一个关键帧提取和相似度计算的方法，针对视频中病灶集中的区域加强时序理解。对无病灶内容也就是无关信息进行丢弃，压缩无关信息，初步筛选粗粒度标签下的超声视频，在时序上重点关注病灶区域。避免视频中不存在病灶信息的帧对前列腺癌肿瘤分期分类器训练的干扰，同时将分割大模型的知识迁移到分类器中。

1. ：现有模型未充分考虑多医生采样之间的个体差异、诊断风格和数据质量波动，导致验证结果不稳定，无法满足医工结合应用中对准确性和鲁棒性的高要求。

为了解决该问题，本发明的大模型先验知识压缩充分考虑了样本之间的差距，对无关内容进行压缩，让所有样本中的重要内容得到了突出。

**4、本发明技术方案的详细阐述，应该结合流程图、原理框图、电路图、时序图进行说明**（所有英文缩写都应有中文注释；**所有附图都应该有详细的文字描述，以别人不看附图即可明白技术方案为准**；同时附图中的关键词或方框图中的注释都尽量用中文；方法专利都应该提供流程图，并提供相关的系统装置）。

技术问题

本发明涉及表征学习领域，具体涉及视频图像文本对齐学习方法。

技术方案

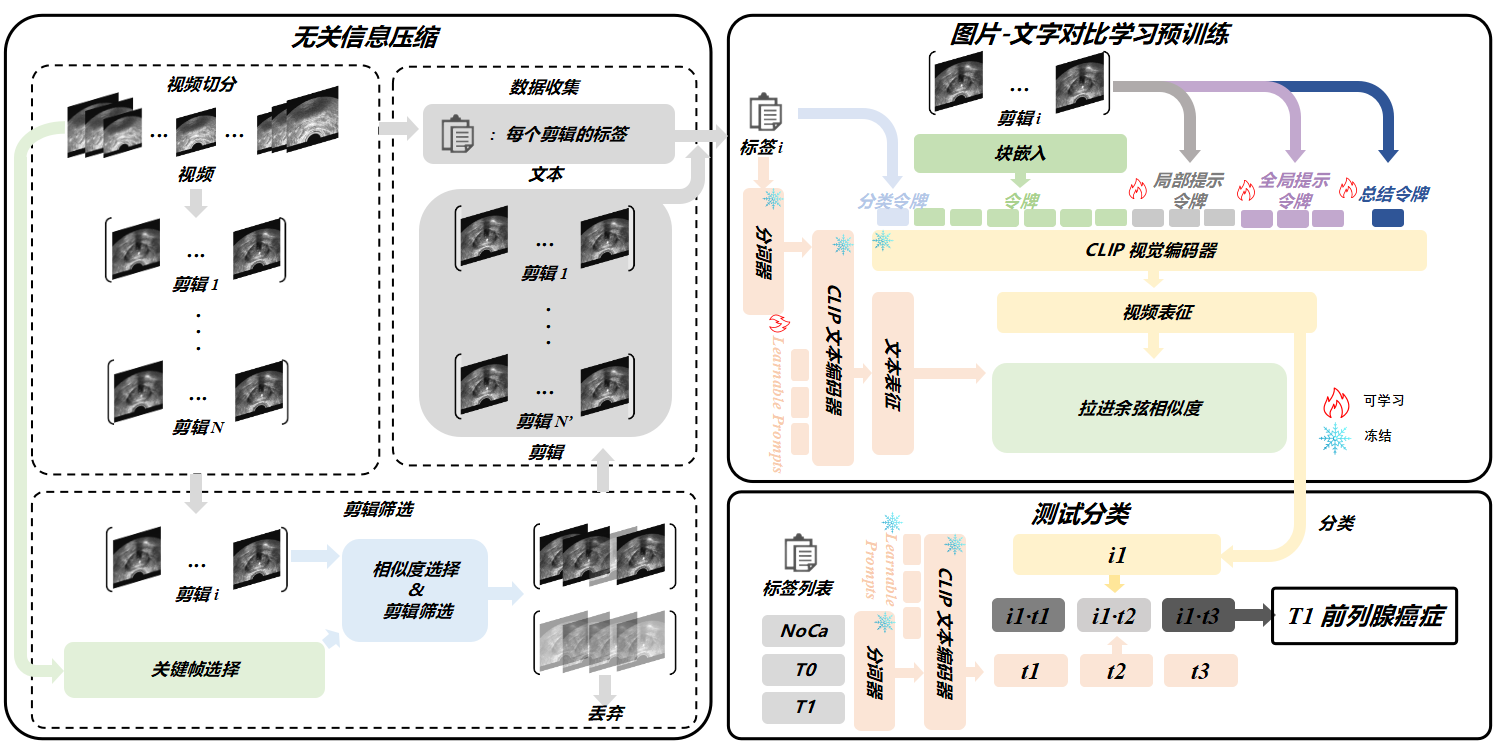


图3.本发明流程框架

一种基于大模型的前列腺癌诊断方法如图3所示。包含以下步骤：

步骤1 数据预处理

将视频进过预训练的大模型MedSAM2，得到分割结果，对有分割结果的视频进行滑动窗口采样，获得一定组数的数据。

步骤2 计算关键帧和相似度，压缩无关帧

通过对整个视频的分割结果的解析，用聚类的方法选择出n类，每一类的中心帧选择为本类的关键帧，针对几帧的病灶大小排序选择唯一关键帧。再通过每一帧和关键帧之间pixel和IOU和帧之间距离计算出相对相似度。针对无关信息，在模型训练的时候需要执行丢弃和重新标注。针对全部帧无分割内容的帧或者全部帧相似度低于阈值的窗口，执行丢弃。

步骤3 文本和图像对齐预训练

针对视频我们对每一帧生成图片令牌（Frames Token），对视频片段生成总结令牌（Summary Token）。经过对比语言-图像预训练（Contrastive Language-Image Pre-Training，CLIP）的视觉编码器（Vision Encoder）之后作为视觉表征；在标签方面，使用可学习的提示（Prompts）生成，经过文本编码器（Text Encoder）之后作为文本表征最后计算相似度。

步骤4 测试分类

通过预训练我们获得的视觉编码器和文本编码器。我们在测试时将前列腺的三个标签都做文本编码获得文本表征，将视频经过视觉编码器之后两者计算余弦相似度，相似度最大的座位最后的测试分类结果。

技术效果

与现有技术相比，本发明具有以下有益效果：

本发明针对其他同领域不同任务的工作，其优越性总结如下：

1. 筛选粗粒度标签机制：对于主要数据的粗标签，本发明提出了一个置信度机制，在大模型的分割下，初步筛选粗粒度标签下的超声视频，在时序上重点关注病灶区域。
2. 无关数据压缩机制：本发明提出了一个基于大模型的前处理机制，在数据端压缩了前列腺超声影像扫描中存在大量无关信息对模型的影响。

附图说明

图1.基于图片的CLIP训练流程框架

图2.基于视频的CLIP训练流程框架

图3.本发明的框架流程

图4.Num\_frames间隔采样精确率召回率

图5.滑动窗口采样精确率召回率

图6.前处理（改label）+滑动窗口采样精确率召回率

图7.前处理（丢弃）+滑动窗口 采样精确率召回率

具体实施方式

下面结合附图和具体实施例对本发明进行详细说明。本实施例以本发明技术方案为前提进行实施，给出了详细的实施方式和具体的操作过程，但本发明的保护范围不限于下述的实施例。

一种基于大模型的前列腺癌诊断方法包括如下具体步骤：

步骤1：数据预处理

针对有T帧的视频V，我们采用步长为2的滑动窗口采样，滑动窗口长度num\_frames 设定为8，这样就能取样出(T-8)//2个视频的片段，此时经过预训练的MedSAM2分割模型，获得一个含有感兴趣区域的分割图像。

步骤2：计算关键帧和相似度，压缩无关帧

从数据的角度出发，超声影像扫描获得的数据存在极强连续性，即有效内容（含病灶内容）相对集中，无效内容（不含病灶内容）也是连续成片段的。因此针对帧之间距离和病灶分割内容我们使用聚类方法进行聚合。对第i帧，我们定义局部密度为

其中dij为第i帧对第j帧的距离，定义为

其中IoUij为两帧之间的分割内容的重叠率，T为整个视频的帧数，Idxij为第i帧和第j帧之间帧索引差的绝对值

再计算每个数据点到相对密度更高点的最小距离

根据每一帧的最小距离选择高密度中心，计算每个点的密度乘积可获得密度乘积的阈值

用来筛选密度中心{}，在密度中心被确定之后进行对密度中心对应的分割图片，通过分割内容最大的高密度中心确定唯一关键帧Fkey\_frame

随后计算每一帧和关键帧Fkey\_frame的相似度Simi

在一开始采样的视频片段找到其对应的相似度

如果一整个视频片段的相似度都低于相似度阈值，那么这个视频片段就会被丢弃，不加入使用。这是由前列腺超声影像数据形式决定的，这样的超声数据采样时一开始和结束的时候有大量无关帧，没有病灶内容，通过前处理压缩掉这些干扰信息。

步骤3：文本和图像对齐预训练

对于每一个分割出来的视频片段，根据VIT的要求，对每一帧打成的N个包，其中，这些向量被扁平化成一组向量，使用一个线性层输出一个D维的向量。加入了一个额外的分类令牌，最后每一帧输出的向量为

经过了L层Video Encoder，代表第i层Video Encoder

最后的视觉特征通过最后一层Encoder之后，经过一个线性层作为输出。

其中代表最后一层编码器层，视频的特征v被表示为

其中V1...Vt代表t长度的片段，这样提取出了视频的表征。我们的改进在最后一步中，在计算之前我们有如下的Token增加

**总结令牌（Summary Token）:**

同时我们对整段视频的特征进行一个总结，叫做总结令牌（Summary Token），对于第i层的Summary Token 我们将每一帧图像在第i-1层视频编码器中的分类令牌取出为

通过一个线性投影，再自执行一次LN层正则化后MHSA多头自注意力再加上自身，即

**全局提示令牌（Global Prompt Tokens）:**

为了让模型获得学习数据分布的能力，随机初始化一队可学习的向量

**本地提示令牌（Local Prompt Tokens）:**

帧等级的提示令牌 也是随机初始化的可学习向量，帧等级本地提示令牌利用了分类令牌

增加完了这些令牌，我们在最后一层编码器中，将，添加到中计算，其中FSA是预训练的自注意力机制

然后将添加的，去除之后，对计算一个前馈神经网络FFN

这样获得的帧表示用来计算帧视频输出

最后池化为视频表征

对于文本的Encoder编码器，我们使用预训练的BERT模型，模型一共有12层，每一层都由MHSA多头自注意力和FFN前馈神经网络组成，设（i=1...12）为第i层文本编码器的输出，那么有

最后在i=12时，获得文本表征，采用提示学习（Prompt Learning）的方法作为文本输入，而不是手工设计的特征比如“这是一个{label}的视频”

总而言之。我们输入视频V和文本C，经过文本编码器和视频编码器后获得视频表征和文本表征如下

对我们提取的v和c特征，定义余弦损失，我们使正确的v，c对最大化，并且最小化其他错误的v,c对。

步骤4 ：测试分类

最后测试分类时，针对给入的剪辑V和标签列表C，我们通过之前预训练好的视觉编码器和文本编码器，获得剪辑V的视频表征v和文本C的文本表征c。计算每一个标签的文本表征和v的余弦相似度，最大余弦相似度做为分类结果。

1. **本发明的关键点和欲保护点是什么？**（对于上一部分给出的详细的完整技术方案，在本部分是提炼出技术方案的关键创新点，列出1、2、3...，以提醒代理人注意，便于专利代理人撰写权利要求书）

无关数据压缩机制：本发明提出了一种新颖的关键帧提取方法，通过大语言模型对视频进行分割并选择具有代表性的关键帧，优化了视频数据的采样过程，避免了无关信息对模型训练的干扰。本发明利用滑动窗口和聚类方法来压缩视频中无关信息，减少了无病灶帧对诊断准确性的影响，从而提高了前列腺癌超声影像的诊断效果。

**6、用推理方式推导出本发明的优点**（务必与第3部分的现有技术的缺点相对应），可以对应3部分所要解决的技术问题或发明目的来描述。

1. 提升了超声视频诊断准确性：相比传统方法需要医生逐帧标注，采用大语言模型进行关键帧提取和相似度计算，不仅减少了医生的工作负担，还有效提升了模型的诊断能力，尤其在处理复杂病灶区域时，能更好地捕捉病灶特征。压缩无关数据减少干扰：通过视频数据的预处理和无关数据压缩机制，避免了超声影像中的无关帧对模型训练的干扰，确保模型训练过程中聚焦于前列腺癌相关的有效信息。
2. 解决了多医生采样间差异问题：本发明充分考虑了不同医生之间的诊断差异及数据质量波动，通过大语言模型的先验知识压缩和自适应的提示学习机制，增强了模型的泛化能力和鲁棒性，使得诊断结果更加稳定、可靠。

**7、针对4中的技术方案，是否还有别的替代方案同样能完成发明目的？**

无

**8、其他有助于专利代理人理解本技术的资料**（给代理人提供更多的信息，可以有助于代理人更好更快的完成申请文件）

无

1. **案例（实施例）**

数据集：

实验在上海第十人民医院的数据集上进行训练和验证，数据统计如下表1所示

表1.本专利数据集统计信息

|  |  |  |
| --- | --- | --- |
| 数据 | 上海第十人民医院（训练） | 上海第十人民医院（验证） |
| 病例数量 | 640 | 160 |
| 无病（Noca）数量 | 200 | 50 |
| 癌症初期（T0期）数量 | 230 | 58 |
| 癌症晚期（T1期）数量 | 210 | 52 |

方法应用：

本实验采用步长为2的滑动窗口采样，将训练批次大小（Batch Size）设置为 64，整个过程中将学习率（Learning Rate,LR）设置为4e-4，在两块Tesla V100显卡上训练。

效果对比

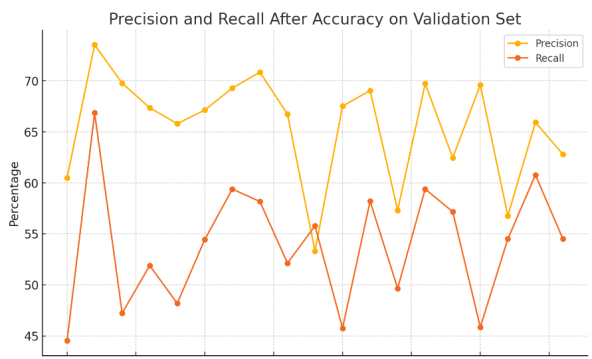
在实验效果对比上，由于医学场景和任务的特殊性，我们将效果对比在工作的消融中。针对采样方法和前处理的方式进行实验。

在采样方法上，对于Num\_frames间隔的采样，即针对T帧视频，采样出T//Num\_frames的片段输入后续的Clip过程。这样的采样方法导致数据的过稀疏，从而导致训练不收敛。然后采用滑动窗口的采样方法，即对于T帧视频，如果步长为stride为s，那么会采样处（T-Num\_frames）//s的片段，经过我们的消融实验，发现stride为2时候实验效果最好。

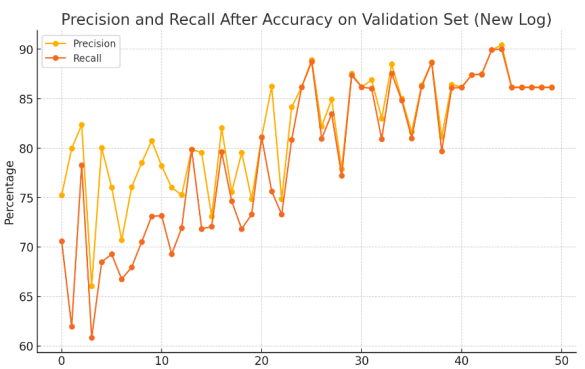
在前处理的具体方式上，针对医生对数据的描述，我们打算采用两种方法：将无关信息帧的label从T0期癌症和T1期癌症修改为良性标签，将无关信息帧片段直接丢弃。从实验结果来看，将无关信息帧直接进行丢弃效果更好。效果如下表2和图所示：

表2.前处理方式实验效果表

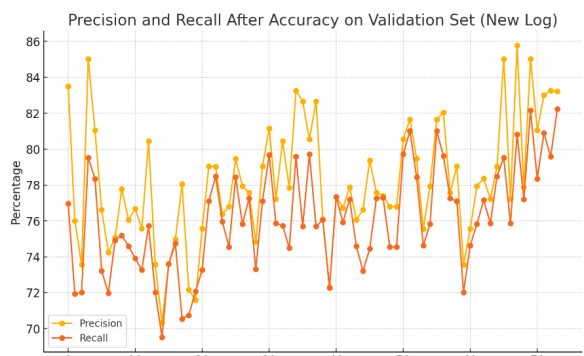
|  |  |  |
| --- | --- | --- |
| 改进方式 | ACC | Pre |
| Num\_frames间隔采样clip | 不收敛 | 不收敛 |
| 滑动窗口采样clip | 90.43% | 90.03% |
| 处理（改label）+ 滑动窗口采样clip | 85.77% | 82.24% |
| 前处理（丢弃）+ 滑动窗口采样clip | **91.55%** | **91.12%** |



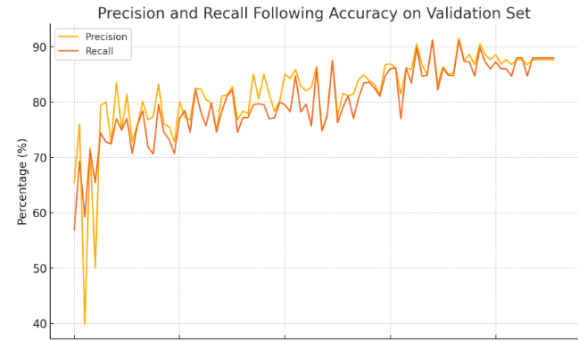
**图4.Num\_frames间隔采样精确率召回率**



**图5.滑动窗口采样精确率召回率**



**图6.前处理（改label）+滑动窗口采样精确率召回率**



**图7.前处理（丢弃）+滑动窗口 采样精确率召回率**

上述描述仅是对本申请较佳实施例的描述，并非是对本申请范围的任何限定。任何熟悉该领域的普通技术人员根据上述揭示的技术内容做出的任何变更或修饰均应当视为等同的有效实施例，均属于本申请技术方案保护的范围