arXiv:2301.02933v1 [eess.IV] 7 Jan 2023

前列腺癌中的弱监督联合全切片分割与分类

Pushpak Pati[†1*]、Guillaume Jaume[†2, 3, 4, 5]、Zeineb Ayadi[8]、Kevin Thandiackal[1,6]、Behzad Bozorgtabar[7]、Maria Gabrani[1]、Orcun Goksel[6,9]

[1]*IBM* 欧洲研究院，瑞士[2]病理学系，布莱根妇女医院，哈佛医学院，美国[3]病理学系，麻省总医院，哈佛医学院，美国[4]癌症项目，哈佛大学与麻省理工学院布罗德研究所，美国[5]数据科学项目，丹娜-法伯/哈佛癌症中心，美国[6]计算机辅助医学应用，苏黎世联邦理工学院，瑞士[7]信号处理实验室*5*，洛桑联邦理工学院，瑞士[8]洛桑联邦理工学院，瑞士[9]信息技术系，乌普萨拉大学，瑞典

摘要—**组织学诊断相关区域的分割与自动识别可为病理学家提供有价值的帮助。然而，由于全切片图像（WSI）的像素级标注既耗时又昂贵，难以获取，分割方法因此受到制约。为解决这一问题，研究人员开发了弱监督方法，以利用图像级别可直接获得的标注。然而，据我们所知，这些技术均未针对全切片图像进行适配。本文提出 WHOLESIGHT，这是一种弱监督方法，可同时对任意形状和尺寸的全切片图像进行分割和分类。具体而言，WHOLESIGHT 首先构建全切片图像的组织图表示，其中节点和边分别描述组织区域及其相互作用。在训练过程中，图分类头对全切片图像进行分类，并通过事后特征归因生成节点级伪标签。随后，利用这些伪标签训练节点分类头，以实现全切片图像分割。在测试过程中，两个分类头同时对输入的全切片图像进行类别预测和分割。我们在三个公开前列腺癌全切片图像数据集上评估了 WHOLESIGHT。我们的方法在所有数据集上均取得了最先进的弱监督分割性能，同时在分类方面达到优于或可比于最先进的弱监督全切片图像分类方法的表现。此外，我们还从分割与分类性能、不确定性估计和模型校准等方面量化了该方法的泛化能力。**

关键词-**计算病理学、全切片图像分割、弱监督学习、弱监督分类、弱监督分割**

一、引言

前列腺癌是美国男性中第二常见的癌症， 2022 [1] 登记的新发病例达25万例，导致3.5万人死亡。然而病理学家在癌症患者的诊断和管理中发挥着关键作用，但其人数正逐渐减少。在美国，2007年至2017年间，病理学家人数下降了18%，导致平均工作量增加了42% [2]。此外，泌尿病理学的实践也面临着诸多挑战 [3]。尽管前列腺癌分级的诊断标准已经确立 [4]，但诊断范围内连续变化的表型特征仍会造成判定差异，观察者内和观察者间变异显著 [5], [6]。人工检查切片同样繁琐且耗时，若能实现自动化和标准化，将大有裨益。这些因素说明，有必要开发计算机辅助诊断（CAD）工具，以实现诊断工作流程的自动化。

为此，研究人员提出了多种基于人工智能（AI）的计算机辅助诊断（CAD）工具，包括细胞核分割与分类 [7]–[9]，腺体分割 [9]–[11]，以及肿瘤检测 [12], [13]。尽管这些工具表现出色，但通常需要在大型数据集上进行针对特定任务和组织的标注；这类标注既繁琐又耗时，且往往难以获取。为减少标注需求，研究人员提出了不同的方法，尤其是基于多实例学习（MIL）框架的弱监督方法，用于自动分类全切片图像（WSI） [14], [15]。

尽管分类很有用，但它在诊断过程中辅助病理学家关注重点方面仍然作用有限。在这种情况下，语义分割方法更为可取，因为它们能够生成组织成分的像素级描绘，从而突出诊断相关区域。这样的可视化有助于增强病理学家与CAD工具之间的信任。此外，分类器还可以利用识别出的区域来改善患者诊断。然而，语义

*通讯作者：Pushpak Pati。电子邮件：pus@zurich.ibm.com

分割通常需要像素级标签，因此其标注需求高于分类任务。基于这一原因，开发弱监督语义分割（WSSS）方法似乎是最合适的应对方案。

尽管 弱监督语义分割 在自然图像上已取得成功，但应用于组织病理图像时会面临各种挑战， [16], 因为它们：(1) 包含具有较大类内差异的细粒度对象 [17]; (2) 组织学成分之间通常存在模糊边界 [18]; (3) 图像可能达到数十亿像素，且组织尺寸不一。然而，已有一些弱监督语义分割 方法被提出用于各种组织学任务。由 [19]–[25]执行弱监督语义分割的图块级方法受限于对图块级标注的需求，以及无法执行具有全局上下文的 全切片图像分割。尽管 [26], [27]将尺度扩展到更大的图块，但在处理全切片图像时会带来很高的计算复杂度和内存需求。由 [25], [26] 提出的方法要求精确的 图块标注来进行模型训练，即，也就是对图块中的每种病变类型进行精确标注，这要求病理学家标注超出标准临床需求的图像。另一方面，近期的全切片图像分类方法利用注意力机制或特征归因来突出显著区域 [14]。尽管这些区域对于视觉评估具有信息价值，但对于准确勾勒相关区域而言，它们并不充分、不完整且较为模糊。此外，生成细粒度显著性图需要密集重叠的图块预测，这在处理全切片图像时计算成本高昂。

鉴于上述WSSS 方法的局限性，我们提出了 WholeSIGHT，即“Whole-Slide SegmentatIon using Graphs for HisTopathology”（利用图进行组织病理学全切片分割），该方法利用WSI级标签，无需任何特定任务假设或后处理，即可同时对任意大小的组织病理图像进行分割和分类。形式上，WholeSIGHT将图像转换为基于超像素的组织图（TG），并将分割问题视为节点分类任务。WholeSIGHT同时结合局部和全局组织微环境来执行上下文分割，主要与基于像素间关系的 WSSS [28]方法一致。总而言之，我们的贡献如下：

- WholeSIGHT是一种新颖的基于图的弱监督方法，可利用现成的WSIs 级标注，对WSI进行联合分割和分类。
- 在3个前列腺癌数据集上对WholeSIGHT进行了全面评估，用于Gleason模式分割和Gleason分级，并与最先进的WSI级弱监督方法进行了基准比较。
- 全面量化 WHOLESIGHT 在域内和域外队列中的泛化能力，评估其分割和分类性能、神经网络预测的不确定性估计与校准。

本研究的初步版本曾以 [29]为题发表。本文的主要扩展包括：（1）改进了 WholeSIGHT 方法的模型架构，并实现了节点标签的自动合成；（2）在大规模全切片图像队列上进行了广泛评估WSIs （约 100×）；（3）开展了泛化评估。

II. 相关工作

A. 弱监督组织病理学图像分类

全切片图像的弱监督分类全切片图像主要围绕多实例学习展开。在多实例学习中，全切片图像首先被分解为一个由图像块组成的“包”，并由神经编码器编码，例如，使用卷积神经网络（CNN）。然后，聚合器汇集图像块嵌入，生成切片级表示，再通过神经预测器映射为聚合器可基于注意力机制，对每个图像块的重要性进行加权，例如 [14], [30], ；也可如近期提出的方法所示，采用 Transformer 的形式 [31], [32] ，或采用图神经网络（GNN） [33], ，从而能够建模图像块间依赖关系和全局上下文。另一种方式是利用全切片图像的多尺度表示来建模上下文，例如使用多放大倍率的图像块嵌入 [23], [34] ，或如 [35]，[36]所提出的，通过学习自动选择重要区域。尽管这些方法取得了成功，但无法直接扩展到语义分割。

B. 弱监督组织病理学图像分割

组织病理学中的弱监督语义分割方法可按监督（或标注）类型（如点标注、涂鸦标注或图像级标签）以及应用尺度（如图像块、图块、组织芯片（TMAs）或全切片图像）进行分类。 [37]–[39]利用点标注对组织学图像块中的细胞和细胞核进行分割。 [23], [40] 则分别在图像块级别使用涂鸦标注分割组织和肿瘤区域。这两种方法都采用了 U-Net [41]，其中 [23]利用跨多个放大倍率的同心图像块来纳入相关上下文信息，而 [40] 则修改目标函数，以平衡标注像素的贡献。组织病理学中的大多数弱监督语义分割方法采用图像级监督，且仅限于处理带有图像块标注的数据。[19] 提出了多重聚类实例学习，用于处理滑动图像块，从而同时对结肠组织芯片进行分级和分割。 [21] 训练了一个二分类器来进行像素级预测，随后通过 Softmax函数根据像素标签计算图像级预测。他们优化图像预测，以提升像素预测效果。 [22] 提出了基于多实例学习的标签扩充方法 CAMEL。该方法将图像划分为网格实例，生成实例标签，并将实例标签赋予相应像素，以实现监督分割。 [26]提出了 HistoSegNet，该方法训练卷积神经网络来预测图块中的组织类型，并利用特征归因推导

像素级预测。该方法还采用了一系列专门的后处理步骤来优化预测结果。 [24] 使用前景比例作为弱标签，并结合全卷积网络和图卷积网络进行组织分割。 [25] 提出了一种基于特征归因的模型来生成伪标签，随后采用多层伪监督网络对组织类型进行分割。这些方法的主要局限在于无法弱监督语义分割在全切片图像上仅凭全切片图像标签实现。为了在弱监督语义分割中突破图像块级别的限制， [27] 提出了 WeGleNet，可扩展至组织芯片。该方法包含一个分割层和一个全局聚合层，用于在训练期间对图像进行分类，并在推理期间对像素级 Softmax激活值进行上采样，以实现图像分割。然而，该方法无法精确描绘病变边界，也无法突出显示多个病变区域。此外，精细分割需要处理大量密集重叠的图像块，且无法扩展至全切片图像。相比之下，我们的 WHOLESIGHT 可利用弱监督语义分割，通过图像级监督实现分割，同时高效扩展至任意尺寸的全切片图像。

C. 组织病理学中的泛化量化

组织病理学中CAD 工具的泛化能力会受到域级偏差的影响。这些偏差由多种因素引起，例如不同的染色方案、制造设备、材料以及具有各自颜色响应的扫描设备 [42]。尽管人们希望获得能够抵御域偏移的通用工具，但在深度学习（DL）中对域偏移进行建模和检测颇具挑战性。尽管如此，已有多种方法被提出，通过数据级和模型级适应来减少此类域偏移。

数据级适应可通过染色归一化 [43]–[46]，颜色增强 [47]，[48]，或染色不变特征学习 [49]，[50]来实现。模型级适应通常通过域对抗训练 [51]–[54],实现，该方法利用目标域的无标签数据以及源域数据进行建模。

然而，上述数据级和模型级适应方法并未全面评估其训练所得深度学习模型在任务表现之外的泛化能力。在这种情况下，准确的不确定性估计和模型校准对于了解何时可以信任模型至关重要——对于经常给出过度自信预测的神经网络而言，这是一项众所周知的挑战 [55]–[57]。据我们所知，计算病理学在这些方向上的研究十分有限，仍有待探索。

III. 方法学

本节介绍用于实现组织病理图像可扩展全视野弱监督语义分割的 W弱监督语义分割。首先，我们将一张全切片图像转换为组织图表示，其中图中的节点和边分别表示组织区域及其相互作用（第III-B节）。接下来，使用图神经网络对节点嵌入进行上下文建模，以表征组织区域（第III-C节），然后通过图分类头进行格里森分级（第III-D节）。最后，我们通过特征归因和节点选择策略生成节点级伪标签，并以此训练节点分类头。节点头输出分割掩膜，并为像素分配 Gleason 模式（第III-E节）。方法概述如图1所示。

A. 符号与预备知识

我们将图 $G \in \mathcal{G}$ 定义为 $(V_G,\ E_G,\ H)$，其中 V_G 和 E_G 分别表示节点集和边集，$H \in \mathbb{R}^{|V| \times d}$ 表示 d 维节点特征（或在节点层级记为 $H_{v,} := h(v) \in \mathbb{R}^d$），$\mathcal{G}$ 是图集。节点 $v \in V_G$ 的邻域记为 $\mathcal{N}(v) := \{u \in V_G \mid (v,\ u) \in E_G \vee (u,\ v) \in E_G\}$。我们用 $|.|$表示集合的基数，例如，$|\mathcal{N}(v)|$ 表示 v的邻居数量。

图神经网络 [58] 是一类从图结构数据中学习的神经网络。具体而言，图神经网络采用两步流程，通过纳入邻域节点信息来融入节点特征的上下文。首先，在 A聚合步骤中，对于每个节点 $v \in V_G$，通过可微且置换不变的函数$\mathcal{N}(v)$ 聚合相邻节点特征。接着，在 U更新步骤中，对 v 的当前特征和 $\mathcal{N}(v)$ 的聚合特征进行可微运算，以更新 v的特征。此流程重复 T 次，其中 T 表示 GNN图神经网络 层数。

在这项工作中，我们采用图同构网络（GIN） [59],其中，AGGREGATE 步骤是一个均值运算符，而 UPDATE 步骤包含一个多层感知机（MLP）。形式上，一个GIN 层定义如下:

$$h^{(t+1)}(v) = \mathrm{MLP}\Big(h^{(t)}(v) + \frac{1}{|\mathcal{N}(v)|} \sum_{u \in \mathcal{N}(v)} h^{(t)}(u)\Big) \quad (1)$$

T GIN 层（记为 $\mathcal{F}_\theta$）堆叠起来，使每个 v都能获取最多 T 跳范围内的上下文信息。对于图分类，通过读出步骤（READOUT step）对节点嵌入 $h^T(v)$ $\forall v \in V_G$ 进行池化，得到固定大小的图级嵌入 h_G；例如，*e.g.*，采用均值运算。随后，h_G通过分类器网络 $\mathcal{F}_\phi$映射到目标类别。类似地，对于节点分类，分类器网络 $\mathcal{F}_\psi$可以对 $h^T(v)$ $\forall v \in V_G$ 进行分类。

形式上，分类旨在预测输入 $x \in \mathcal{X}$ 的目标标签 $y \in \mathcal{K}$，其中 $\mathcal{K}$ 和 $\mathcal{X}$ 分别表示类别集合和输入。给定一组样本对 $\{(x_i, y_i)\}_{i=1}^N$，其中 N 为样本数量，且 $(x_i, y_i) \sim p(x, y)$，数据似然可表示为 $p(Y|X, \theta, \phi) = \Pi_{i=1}^N p(y_i|x_i, \theta, \phi)$。通过最大似然估计，或等价地，通过最小化负对数似然（NLL） $-\sum_{i=1}^N \log\ p(y_i|x_i, \theta, \phi)$，可得到最优参数 $\hat{}\hat{}(\theta, \phi)$。对于图分类，样本对表示为 (y_G, G) $y_G \in \mathcal{K}_\mathcal{G}$，$G \in$

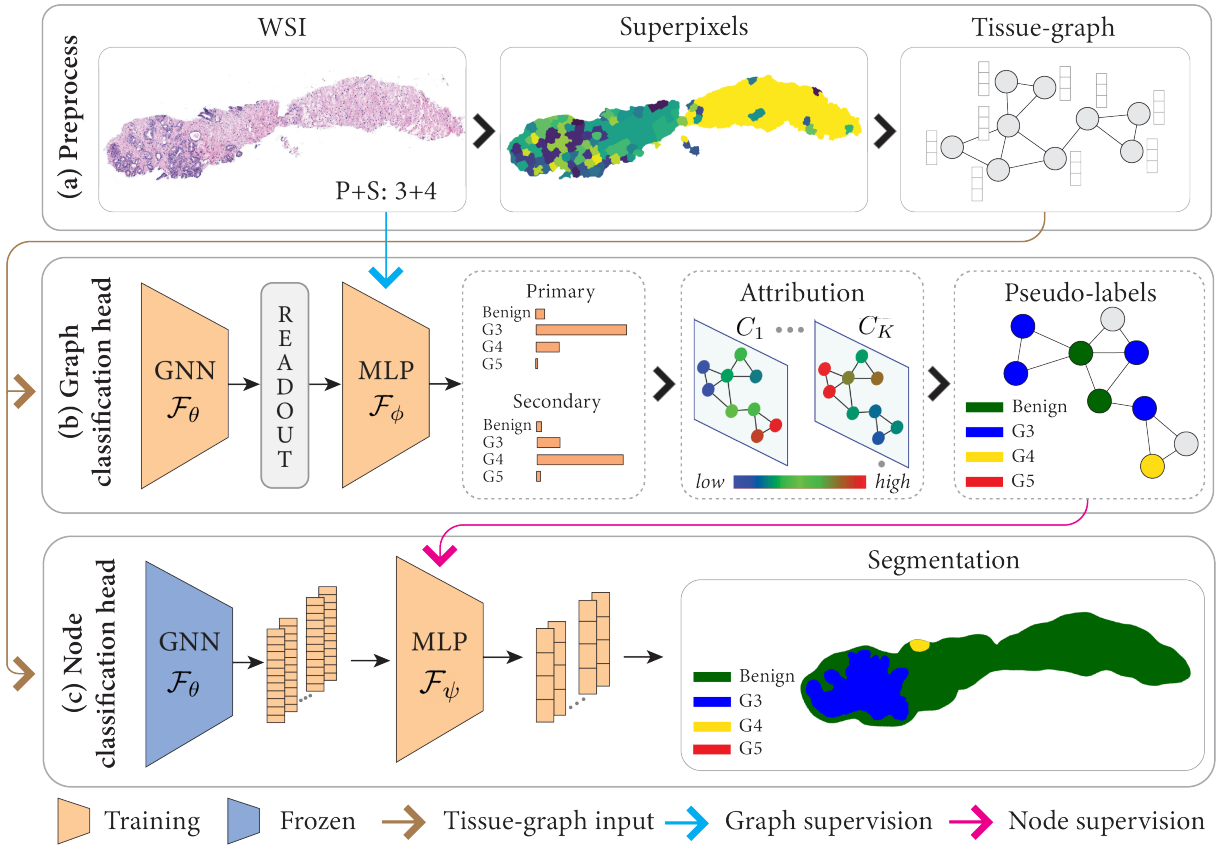

图1：所提出的 WHOLESIGHT 方法概述。(a) 在预处理步骤中，构建组织图 来表示一幅全切片图像，其中通过识别超像素和区域邻接连通性，分别定义节点和边。(b) 图分类头将组织图分类为主要和次要 Gleason 模式。随后，利用特征归因技术和节点选择策略生成节点级伪标签。(c) 节点分类头基于伪标签进行学习以分类节点，从而实现全切片图像分割。

$\mathcal{G}$。对于节点分类，样本对表示为 $(y_V,\ v)$ $y_V \in \mathcal{K}_\mathcal{V}$，$v \in \mathcal{V}$。对于本文的任务，图级和节点级类别集合相同，即 *ie.*，$\mathcal{K} := \mathcal{K}_\mathcal{G} = \mathcal{K}_\mathcal{V}$。

我们进一步引入模型校准 [60]的概念。直观而言，经过校准的模型，其预测结果的概率，即置信度分数，应与模型的实际表现相符。例如，模型以平均 60% 的置信度进行预测的样本，其平均准确率应为 60%。形式上，对于给定的网络，$f: \mathcal{X} \to \mathcal{K}$，以及 $p(X,\ Y)$ 数据与标签的联合分布 $f(x)$ 若相对于 p 满足以下条件，则称其经过校准：$\mathbb{E}_p[Y|f(X) = \beta] = \beta$，$\forall\beta \in [0,1]$。校准可以通过可靠性图 [61]进行可视化。具体而言，数据集中的所有样本都根据其预测置信度分数被分配到各个分箱中。随后，计算每个分箱内样本的模型准确率。将网络性能与分箱后的置信度分数绘制成图，偏离对角线的分箱表示未校准。

B. 预处理与组织图构建

首先，我们使用 [44]提出的方法对输入的 H&E染色图像进行染色归一化。这可以减少组织制备过程中造成的图像间外观差异，即不同的标本制备技术、染色方案、固定特性和成像设备特性 [62], [63]。然后，我们将归一化后的图像转换为组织图(图1(a))，其中，组织图的节点和边分别表示组织区域和组织间相互作用。受 [64], [65], 启发，我们将超像素视为编码组织区域的视觉基元。与矩形图像块相比，超像素能够根据组织的局部同质性更灵活地适应任意形状。同质性约束还限制了超像素跨越多个不同结构并包含不同形态学区域。

组织图 的构建遵循以下步骤 [66], ： (1) 检测超像素以定义节点 V_G，(2) 描述超像素以定义节点特征 H，以及 (3) 构建图拓扑以定义边 E_G。我们

采用两步法在 全切片图像中识别超像素。首先，我们使用简单线性迭代聚类（SLIC） [67]生成过分割的超像素。我们在低放大倍率下进行过分割，以捕捉同质区域，同时在颗粒度与平滑噪声之间取得良好平衡。第二步，根据高放大倍率下的逐通道颜色相似度，对过分割的超像素进行层次合并。颜色相似度通过逐通道的 8区间颜色直方图、均值、标准差、中位数、能量和偏度来量化。合并后得到的组织区域构成组织图的节点。合并后的超像素表示在形态学上有意义的同质区域。此外，合并降低了组织图的节点复杂度，从而使组织图 能够扩展至大型全切片图像 ，并实现对远距离组织区域的上下文化。

我们通过形态学和空间特征来表征组织图节点。考虑到超像素的维度可能不固定，我们采用两步流程来提取形态学特征。首先，从一个超像素中提取大小为 144×144 像素的图像块，将其调整为 224×224 大小，并通过在 ImageNet 上预训练的 MobileNetV2 网络将其编码为 1280 维特征 [68] [69]。超像素级特征通过对图像块级特征取均值得到。接下来，我们通过用图像尺寸对超像素质心进行归一化，为每个节点计算空间特征。归一化确保空间特征不受输入全切片图像尺寸变化的影响。最后，我们通过基于超像素的空间连通性构建区域邻接图拓扑组织图，来定义 [70] 的边。为此，我们假设相邻的组织区域之间存在最强的生物学相互作用，因此应在组织图中相连。

C. 节点嵌入的上下文化

给定一个组织图，我们利用节点上下文信息（见图1（b））学习具有判别性的节点嵌入，即， 组织微环境和组织间相互作用。具体而言，我们使用图同构网络 [59] ，记为 $\mathcal{F}_\theta$。由于图神经网络可处理任意大小且尺寸各异的图，因此能够对以组织图形式表示的组织病理图像进行编码，而无需采用基于图块的处理。由于节点的判别信息取决于其局部子图结构，并且可能存在于图神经网络的不同抽象层级中，我们采用跳跃知识 [71] 策略来利用多层级节点表示。具体而言，经过 T 图同构网络层后的最终节点级嵌入定义如下：

$$h^{(T)}(v) = \mathrm{CONCAT}(h^{(t)}(v),\ \forall t \in \{1, ..., T\}) \quad (2)$$

其中，CONCAT 表示拼接操作。

D. 全切片图像分类

节点特征经过上下文化后，图级分类头使用图级组织图嵌入对其进行分类嵌入 h_G 和图级/图像级监督。为获得固定大小的 h_G，我们使用读出操作对节点嵌入 $h^{(T)}(v)$ $\forall v \in V_G$进行平均。随后，将 h_G 输入多任务分类器，对主要和次要 Gleason 分级进行预测。具体而言，该分类器包含两个多层感知机（MLP），记为$F_\phi = \{F_{\phi_1}, F_{\phi_2}\}$，分别预测全切片图像中的主要模式（即最差的 Gleason模式）和次要模式（即次差的 Gleason模式）。每个多层感知机都解决一个包含 $|\mathcal{K}|$ 种 Gleason模式的多类别问题，即良性、3级、4级和5级。最终的 Gleason 分级由预测的主要模式和次要模式相加得出。$\mathcal{F}_\theta$ 和 $\mathcal{F}_\phi$ 通过最小化加权交叉熵损失进行联合优化，

$$\mathcal{L}_G = \lambda \mathcal{L}_{CE}(y_{G_P}, \hat{y}_{G_P}) + (1-\lambda)\mathcal{L}_{CE}(y_{G_S}, \hat{y}_{G_S}) \quad (3)$$

其中， P 和 S 分别表示真实标注 y_G 和预测$\hat{}y_G$的主要和次要标签， $\lambda \in [0,1]$ 是用于平衡这两项的超参数。此外，在训练期间，我们将类别权重设为 $w = \{ \quad i =: \log(\{1, ..., |\mathcal{K}|\}\} \quad N_i$ $\sum_i N_i$)，其中是训练集中各类别格里森模式的计数全切片图像。这些权重通过为频率较低的类别赋予更高权重，解决格里森分级中的类别不平衡问题。

E. 弱监督语义分割

组织图中的节点 用于识别超像素，即形态学同质组织区域。由于每种 Gleason模式都具有独特的形态学模式，我们假设每个由节点表示的组织区域都包含一种独一无二的Gleason模式。因此，全切片图像分割任务可转化为对组织图中的节点进行分类。在只有图像监督的情况下，节点分类分两步完成。首先，利用图像标签生成伪节点标签，然后使用伪标签训练节点分类器。

伪节点标签：在全切片图像分类之后，使用一种事后特征归因技术来衡量各节点对组织图分类的重要性。具体而言，我们使用 GRAPHGRAD-CAM [72], [73],，这是梯度加权类激活映射 [74]的一种扩展，可用于图神经网络。给定一个图 G，GRAPHGRAD-CAM 会生成按类别划分的节点归因图， A_k、 $\forall k \in \mathcal{K}$。如图1所示，这些图突出显示了对将 G 分类为 $|\mathcal{K}|$的重要性 $\forall v \in V_G$ 。给定 v 对于 $|\mathcal{K}|$的重要性分数，可以假设如果最高重要性分数对应类别 k，则 v 的标签为 $k \in \mathcal{K}$。此时，可以对 A_k、 $\forall k \in \mathcal{K}$ 执行 argmax操作来分类节点。然而，这种节点标注可能并非最优，因为，

- 有些节点仅起到微弱作用，对图的分类贡献很小，因此其重要性分数较低 $\forall k \in \mathcal{K}$ 。然而，对某个节点的重要性分数执行 最大值索引

会贪心地选取得分最高的类别，即使尚未确定该节点的标签 .

- 对某一类别的预测贡献很大的节点，不一定属于该类别。例如，如果一个节点提供了有用的补充信息，用于平局判定或排除另一种类别的可能性，那么它可能具有很高的重要性。形式上，如果节点集合 $V_k \subset V$ 对于类别k具有较高的重要性分数，则无法确保 V_k 的标签为 k。甚至也无法保证 $v \in V_k$ 的标签相同。

- 类别归因图不一定会突出显示属于该类别的所有节点。根据任务的复杂程度，分类器可能只利用某个类别中一部分具有信息量的节点来预测图标签。形式上，如果节点集合 $V_k \subset V$ 对于类别 k具有较高的重要性分数，那么 V_k 可能不包含 $\mathcal{V}_k \subset V$ 中所有实际标签为 k 的节点，即.例如., $V_k \subset \mathcal{V}_k$。

- 鉴于文献中存在多种具有不同底层机制的特征归因技术，它们可能会生成不同的归因图 [73]。因此，对于基于分数的节点分类，单一的归因技术可能并不可靠。

因此，我们计划将通过特征归因突出显示的节点用作伪标签，以训练节点分类器。对于一个具有 Gleason评分 $P+S$ 的图 G $PS \in \mathcal{K}$，我们使用 GRAPHGRAD-CAM 计算节点重要性分数 I_P 和 I_S $\forall v \in V_G$。由于 GRAPHGRAD-CAM 得出的分数没有界限，我们使用最小-最大方法对分数进行归一化。然后，我们选出超过阈值 t 的排名前 n% 的节点，记为 V_P 和 V_S，其中 n 和 t 是在训练期间调整的超参数。该方法会选出对下游节点分类最具信息量的节点。对于节点 $v \in V_P$ 和 $v \in V_S$，我们使用 arg max($I_P(v)$ $I_S(v)$) 来确保 $V_P \cap V_S = \emptyset$。最后，将得分最高的类别指定为节点的伪标签 $y_{\tilde{V}}$。对数据集中的所有组织图重复这一过程，即可得到伪标签 $Y_{\tilde{V}}$。

节点分类： $Y_{\tilde{V}}$ 用于训练节点分类头，如图1所示。具体而言，对于图 G，我们使用 $\mathcal{F}_{\hat{\theta}}$获取节点嵌入 $h^{(T)}(v)$ $\forall v \in V_G$ ˆ，其中 θ 是图神经网络的参数。在节点分类期间， $\mathcal{F}_{\hat{\theta}}$被冻结，因此分割和分类使用相同的图神经网络 骨干网络，从而减少可训练参数的数量。$h^{(T)}(v)$ $\forall v$ 由多层感知机 $\mathcal{F}_{\psi}$处理，以预测 $Y_{\tilde{V}}$。通过优化加权多类别交叉熵目标来训练$\mathcal{F}_{\psi}$ 。与图分类类似，类别权重定义为 $w := \{\log(\frac{\sum_i N_i}{N_i})$ $i = \{1, ..., |\mathcal{K}|\}\}$，其中 N_i 是类别 i的已标注节点数量。逐节点预测类别标签用于获得最终的分割预测。值得注意的是，与 [26],不同，WHOLESIGHT不包含任何定制的后处理因此适用于各种组织和分割任务。

值得注意的是，对于图 G，图-和节点分类头分别处理互补任务，即图级别和节点级别的任务。因此，在ˆ完成 $\mathcal{F}_{\psi}$ 的训练后，我们解冻 $\mathcal{F}_{\theta}$，并以较小的学习率联合微调 θ ˆ和 ψ 。这些任务的互补性提供了额外的信息信号，进一步提升 WHOLESIGHT 的分割和分类性能。

IV. 实验

A. 数据集

我们在三个数据集上评估所提出的方法，这些数据集包含用于 Gleason模式分割和格里森分级的全切片前列腺癌针芯活检。Gleason模式包括3级（G3）——中度分化细胞核和形成不良的筛状腺体、4级（G4）——分化不良的细胞核和不规则团块，以及5级（G5）——分化程度更低的细胞核，且缺乏腺体或仅偶见腺体。正常腺体和非上皮组织标记为良性（B）。Gleason 分级描述全切片图像中最严重（主要，P）和次严重（次要，S）的 Gleason模式。全切片图像。数据集详情如下：

Sicap 数据集： 该数据集 [75] 包含 18,783 个尺寸为 512×512 的图像块，具有完整的像素级标注和 155 张全切片图像的切片级格里森分级，来自 95 名患者。原始切片和掩膜通过拼接图像块重建而成。全切片图像由 Ventana iScan Coreo 扫描仪以 40× 放大倍率扫描，并降采样至 10× 放大倍率。这些切片由西班牙瓦伦西亚 Clínico 医院的专家泌尿生殖病理学家进行标注。

Radboud 数据集： [76] 包含来自荷兰 Radboud大学医学中心 1,243 名患者的 5,759 份针吸活检。切片使用 3D Histech Panoramic Flash II 250 扫描仪以 20× 放大倍率扫描（分辨率为 0.24μ米/像素），并降采样至 10×。标注包括 全切片图像的 Gleason 分级，以及带噪声的像素级 Gleason 模式掩膜；这些数据作为前列腺癌分级评估（PANDA）挑战赛的一部分发布 [77]。我们使用标准图像处理技术清理用于分割的掩膜，即，，上下文噪声去除、孔洞填充和边缘平滑。由于缺少带有像素级前列腺癌标注的大型公开全切片图像数据集，我们使用该数据集开发并评估我们的方法。

*Karolinska*数据集：该数据集 [78] 包含来自瑞典斯德哥尔摩多家医院的 1,222 名患者的 5,662 份粗针活检样本。切片使用 Hamamatsu C9600-12 和 Aperio Scan Scope AT2扫描仪以 20× 放大倍率扫描，像素分辨率分别为 0.45202μ米和 0.5032μ米。活检样本由一位泌尿病理学专家进行格里森分级标注。

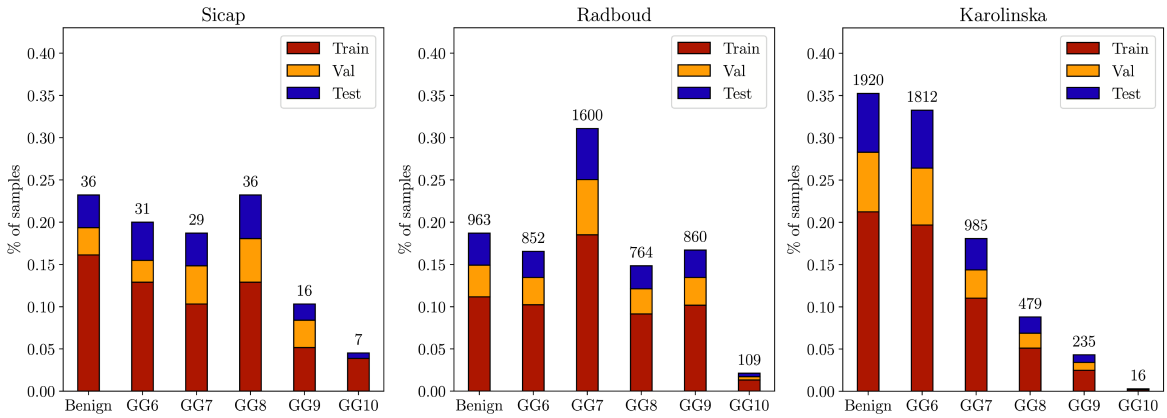

图 2：按 Gleason 分级　　展示 Karolinska、Radboud 和 Sicap 数据集中训练集、验证集和测试集的数据分布　　ets.

每个数据集均按 Gleason 分级，以 60%、20% 和 20% 的比例划分为训练集、验证集和测试集，并采用随机分层，确保各划分中类别所占百分比保持一致。数据集分布和划分情况见图2，其中突显了类别级不平衡。

B. 实现与评估

我们使用 PyTorch 实现了 WholeSIGHT [79],DGL [80], 和 Histocartography [81], , 并在 NVIDIA Tesla P100 GPU 和 POWER9 CPU 上进行了实验。

为开发 WholeSIGHT 网络， $\mathcal{F}_\theta$、 $\mathcal{F}_\phi$和$\mathcal{F}_\psi$ 均通过优化各自的超参数进行设计。首先， $\mathcal{F}_\theta$ 和 $\mathcal{F}_\phi$ 使用图像级标签进行训练，然后创建伪节点标签来训练 $\mathcal{F}_\psi$ ，以生成分割输出。将 图同构网络 层的数量 $\mathcal{F}_\theta$ 优化为 $\{3,4,5\}$，其中 Update 函数定义为一个具有 64 个隐藏单元和 ReLU 激活的两层 多层感知机。 $\mathcal{F}_\phi$ 包含两个用于分类主要和次要Gleason 分级的头部，每个头部均由一个具有 128 个隐藏单元和 ReLU 激活的两层多层感知机组成。 $\mathcal{F}_\psi$ 包含一个具有 128 个隐藏单元和 ReLU 激活的两层多层感知机。

考虑到 Sicap 数据集规模较小，我们采用节点级增强来扩充训练集。具体而言，使用随机节点旋转 $\{90,180,270\}$ 度以及水平和垂直镜像来增强节点。批量大小和学习率分别从 $\{4,8,16\}$ 和 $\{10^{-4},5\times10^{-4},10^{-3}\}$ 中优化确定。属于 $\mathcal{F}_\theta$、 $\mathcal{F}_\phi$ 和 $\mathcal{F}_\psi$ 的 多层感知机 中分别加入了 Dropout 层，丢弃率为 0.2、0.5 和 0.5。针对 $\{5,10,15,20\}$ 中的选择比例和阈值$\{0.5,0.6,0.7\}$提取了伪节点标签。完成超参数调优后，使用不同的网络初始化训练了十个 WHOLESIGHT 模型。根据验证集加权 F1 分数进行模型选择。报告的结果为这十个模型的均值和标准差。

分类指标： 分类性能通过 Gleason 分级的加权F1分数和 ISUP 分级的二次加权 Kappa 分数（κ^2）进行衡量 [82], [83]。ISUP 是一种替代分级系统，与格里森分级的对应关系定义如下：良性 $\rightarrow$ ISUP-0，腺体分级-（3+3） $\rightarrow$ ISUP-1，腺体分级-（3+4） $\rightarrow$ ISUP-2，腺体分级-（4+3） $\rightarrow$ ISUP-3，腺体分级-8 $\rightarrow$ ISUP-4，腺体分级≥9 $\rightarrow$ ISUP-5。 κ^2 该分数反映预测结果与真实标签之间的分歧程度。例如，将等级为 6 的样本预测为等级 10，所受惩罚大于将其预测为等级 7。

分割指标： 分割性能通过 Dice 分数进行衡量。鉴于数据集中 Gleason模式分布不均，我们还报告了各模式的 Dice 分数。

不确定性指标： 按照 [84]，的研究，我们使用 Brier分数 s_{B} （越低越好）和 负对数似然 s_{NLL} （越低越好）来评估一组 N 测试样本的分类不确定性，定义如下：

$$s_{\text{B}} = \frac{1}{N}\sum_{n=1}^{N}\sum_{i=1}^{|\mathcal{K}|}(y_i-\hat{y}_i)^2, \quad s_{\text{NLL}} = -\frac{1}{N}\sum_{n=1}^{N}\sum_{i=1}^{|\mathcal{K}|}p(y_i)\log\hat{p}(y_i) \tag{4}$$

校准指标： 可靠性图直观地展示了模型校准情况。为量化这些观察结果，我们使用了期望校准误差（ ECE）指标 [85], , 该指标计算所有分箱中置信度分数偏差的加权平均值，即。, :

$$c_{\text{ECE}} = \sum_{b=1}^{B}\frac{N_b}{N}|\text{acc}(b)-\text{conf}(b)| \tag{5}$$

其中， n_b 是分箱 b中的样本数，acc(b)和conf(b)分别表示 b中样本的准确率和平均置信度。

C. 基线方法

我们将 WholeSIGHT 与最先进的全切片图像分类方法以及 WHOLESIGHT 的两种变体进行了比较。

FSConv：我们实现了 [75] 提出的用于全切片图像分类的两步法。首先，我们从全切片图像中提取尺寸为 256×256 的图像块，并使用 FSConv+全局最大池化对其进行分类。使用 Gleason 模式掩膜为图像块标注，并选取包含 >90% 单一模式的图像块进行分类器训练。推理时，密集图块预测生成输出分割掩膜。我们在多层感知机上使用全切片图像图像块的 Gleason 分级百分比进行训练，以实现全切片图像的格里森分级。

WHOLESIGHT（图，GRAPHGRAD-CAM）：与 WHOLESIGHT 相比，此基线方法仅包含 $\mathcal{F}_\theta$和 $\mathcal{F}_\phi$。它不生成或使用伪标签，分割输出是通过对按类别划分的 GRAPHGRAD-CAM 归因图取最大值索引 获得的。

WHOLESIGHT（多重，*NC*）：此变体在训练期间同时使用图像级和像素级监督，是 WHOLESIGHT 的上界。由于有像素级标注，$\mathcal{F}_\psi$ 使用真实节点级标签进行训练，而不是生成伪节点标签。该模型由与 W $\mathcal{F}_\theta$、 $\mathcal{F}_\phi$和 $\mathcal{F}_\psi$相同的组件构成，采用 WHOLESIGHT 架构。在此基线方法中，$\mathcal{F}_\theta$、$\mathcal{F}_\phi$和 $\mathcal{F}_\psi$ 通过优化多任务目标进行联合训练，即。*e*.，全切片图像级主要和次要 Gleason 评分预测，以及节点级 Gleason 模式预测。此 WHOLESIGHT 变体是在我们的初步研究中提出的，详见 [29]。

多实例学习（*MIL*）：MIL 方法是全切片图像分类领域的先进方法。具体而言，我们将其与 ABMIL [30], 进行比较。该方法使用注意力机制将图像块嵌入聚合为固定大小的全切片图像嵌入，并将其输入分类器进行格里森分级。我们还纳入了 CLAM [14], ，这是一种基于 ABMIL 构建的方法，通过加入额外的约束来聚类相似的图像块嵌入。我们的实验遵循公开实现*，并进行调整以支持多任务分类。

对于所有基线方法，我们都对超参数进行了全面调优，以在适用时选取最佳学习率和批量大小。随后，使用最优参数从头开始重新训练十个模型。我们报告了每项实验中这十次运行结果的均值和标准差。

D. 弱监督语义分割性能分析

我们通过分别在 Sicap、Radboud 和 Karolinska 上训练和测试 WHOLESIGHT 及其竞争方法，研究其分类和分割性能。

性能分析：表I 展示了 Sicap 上的结果。分析分为两种监督设置，即，完全监督（$\mathcal{C}$）和弱监督（$\mathcal{W}$）。设置-$\mathcal{C}$ 同时使用图像级和像素级标注，而设置-$\mathcal{W}$ 仅使用图像级标签。WHOLESIGHT 达到了 37.6% 的平均 Dice 分数，绝对值显著优于 WHOLESIGHT（图，GRAPHGRAD-CAM） +6.6%。作为上界的 WHOLESIGHT（多重染色，NC）在分割方面较 WHOLESIGHT 取得了显著提升。每类 Dice 分数表明，与出现频率较低的 Gleason模式相比，占大多数组织区域的良性模式具有较高的检出率。在分类任务中，无论是 Gleason 分级加权 F1 还是 ISUP κ^2，WHOLESIGHT 的表现均优于 ABMIL 和 CLAM。值得注意的是，

表II 展示了 Radboud 数据集上的结果。与 WHOLESIGHT（图，GRAPHGRAD-CAM）相比，WHOLESIGHT 的平均 Dice 分数绝对提高了 +4.8%。这证实了伪节点标签有助于实现更优的分割。与 Sicap 数据集上的观察结果类似，良性模式的检出率较高，其次是 G3、G4 和 G5 模式。由于 Radboud 数据集中的 G5 模式多于 Sicap 数据集，我们观察到高等级 Gleason模式的检出率显著提升。在分类任务中，观察结果也与 Sicap 数据集上的一致。

表III展示了 Karolinska 数据集上的结果。由于缺少像素级真实标注，无法计算分割性能。就分类性能而言，WHOLESIGHT（图）优于基线方法。

表I、III 和III中的观察结果表明，联合优化分类和分割目标可提供互补信息，从而提升整体分类性能，即. *e*.，WHOLESIGHT（多重）> WHOLESIGHT > WHOLESIGHT（图）。

E. 泛化：性能、不确定性与校准

我们研究了遵循经修改训练设置的 W全切片图像的泛化能力。具体而言，我们使用 Radboud 和 Karolinska 训练集全切片图像进行模型训练。因此，与在各个数据集上单独训练相比，训练集涵盖了更丰富的样本变异性，以及诊断难度更高的病例。我们分别在 Radboud 和 Karolinska 测试集全切片图像上进行测试，以研究域内性能。此外，我们在整个 Sicap 数据集上进行了测试，该数据集包含域外全切片图像。

性能分析：表IV 比较了 WHOLESIGHT 与竞争基线方法的分类表现。在域内测试集上，就加权F1分数而言，WHOLESIGHT 的表现优于 ABMIL，且优于或相当于 CLAM。ISUP κ^2也呈现出类似的趋势。不过，WHOLESIGHT 的分类结果方差始终低于 ABMIL 和 CLAM。在域外Sicap 数据集上测试时，WHOLESIGHT 的分类表现显著优于 ABMIL 和 CLAM。

*CLAM 公开代码：https://github.com/mahmoodlab/CLAM

表 I：**Sicap** 数据集上的分类和分割结果。使用图像级监督时的最佳表现以**粗体**标出。

Annot.	方法	每类 Dice 良性	3级	4级	5级	平均 Dice	GGwF1	ISUP	κ^2
$\mathcal{C}$	FSConv [75]	65.7±0.5	24.4±1.4	29.0±1.4	8.4±0.6	31.9±0.5		48.7±3.4	50.9±3.4
	WHOLESIGHT（多重染色，NC）	92.5±0.3	35.4±2.3	51.6±2.0	11.1±2.2	47.6±2.1		59.6±4.1	84.6±3.2
$\mathcal{W}$	ABMIL [30]	-	-	-	-	-		50.2±6.3	67.8±5.2
	CLAM [14]	-	-	-	-	-		51.4±5.5	75.2±4.7
	WHOLESIGHT (图, GRAPHGRAD-CAM)	64.4±6.1	23.1±2.0	32.8±6.9	3.7±1.0	31.0±2.7		53.3±5.3	81.9±6.7
	WHOLESIGHT （图 + 伪节点，节点类别。）	**67.6±0.5**	**28.6±0.3**	**49.1±0.4**	**5.0±0.4**	**37.6±0.3**		**58.6±6.2**	**89.0±1.2**

表 II：Radboud 数据集上的分类和分割结果。使用图像级监督时的最佳表现以**粗体**突出显示。

Annot.	方法	每类 Dice 良性	3级	4级	5级	平均 Dice	GGwF1	ISUP	κ^2
$\mathcal{C}$	FSConv [75]	84.3±0.1	53.3±0.5	62.5±0.3	36.9±0.5	59.2±0.1		45.9±1.3	69.4±0.6
	WHOLESIGHT（多重染色，NC）	91.5±0.1	63.9±0.4	66.2±0.3	36.8±1.2	64.6±0.4		68.9±0.9	83.8±0.9
$\mathcal{W}$	ABMIL [30]	-	-	-	-	-		59.3±1.7	79.7±1.2
	CLAM [14]	-	-	-	-	-		60.3±1.6	80.2±1.5
	WHOLESIGHT (图, GRAPHGRAD-CAM)	71.3±2.2	26.9±1.0	24.9±1.2	15.1±0.5	34.6±0.6		66.0±1.0	82.2±0.5
	WHOLESIGHT （图 + 伪节点，节点类别。）	**75.9±0.3**	**32.9±1.0**	**29.1±1.5**	**19.6±0.7**	**39.4±0.3**		**67.9±0.3**	**83.0±0.2**

表 III：Karolinska 数据集上的分类结果。使用图像级监督时的最佳表现以**粗体**标出。

		GGwF1	ISUP	κ^2
$\mathcal{W}$	ABMIL [30]	65.0±2.0	79.1±1.2	
	CLAM [14]	63.6±2.5	77.6±2.0	
	WHOLE视野（图）	**70.5±0.6**	**80.2±0.7**	

表IV：使用 Radboud 和 Karolinska 数据集训练的模型在 Radboud、Karolinska 和 Sicap 数据集上的分类和分割结果。

Annot.	方法	Radboud 平均 Dice	腺体分级	加权F1分数	ISUP	κ^2	Karolinska 腺体分级	加权F1分数	ISUP	κ^2	Sicap 平均 Dice	腺体分级	加权F1分数	ISUP	κ^2
$\mathcal{C}$	FSConv [75]	59.2±0.1	45.9±1.3	69.4±0.6	34.5±1.1	40.1±1.3	49.5±0.4				52.1±2.1	53.8±1.7			
	WHOLESIGHT（多重染色，NC）	64.5±0.3	69.0±1.0	83.6±0.9	71.2±0.7	82.5±1.3	60.0±0.5				65.5±2.5	85.6±2.8			
$\mathcal{W}$	ABMIL [30]	-	57.6±2.3	73.8±2.3	65.5±1.3	77.3±2.8	-				56.4±2.7	75.0±7.5			
	CLAM [14]	-	61.7±2.1	78.6±1.3	69.3±1.3	**82.8±1.0**	-				53.1±3.8	74.6±4.2			
	WHOLESIGHT (图，梯度加权类激活映射)	34.6±0.6	66.0±1.0	82.2±0.5	69.2±0.9	80.3±0.9	30.4±1.0				65.1±2.3	86.1±2.5			
	WHOLESIGHT （图 + 伪节点，节点类别。）	**44.6±0.2**	**66.2±0.1**	**82.9±0.1**	**70.2±0.1**	81.3±0.1	**42.0±0.3**				**65.2±0.1**	**86.6±0.1**			

WHOLESIGHT 在三个数据集上进行最佳格里森分级、ISUP分级以及原发和继发分类的混淆矩阵见图3。可以看出，大多数误分类都接近对角线。大部分混淆发生在腺体分级6 和腺体分级7之间，即腺体分级。，腺体分级(3 + 3)与腺体分级(3 +4）以及腺体分级(4 + 3）。这种模糊性在以下情况中很常见病理学家，如 [86], [87]所示。由于 Radboud 的高级别样本数量更多，因此其高级别格里森分级表现优于 Karolinska。Radboud、Karolinska 和 Sicap 的原发和继发分类加权 F1 分别为 79.3%、81.7%、81.6% 和 62.5 %、69.7%、64.5%。这表明识别次要格里森模式更具挑战性。表IV

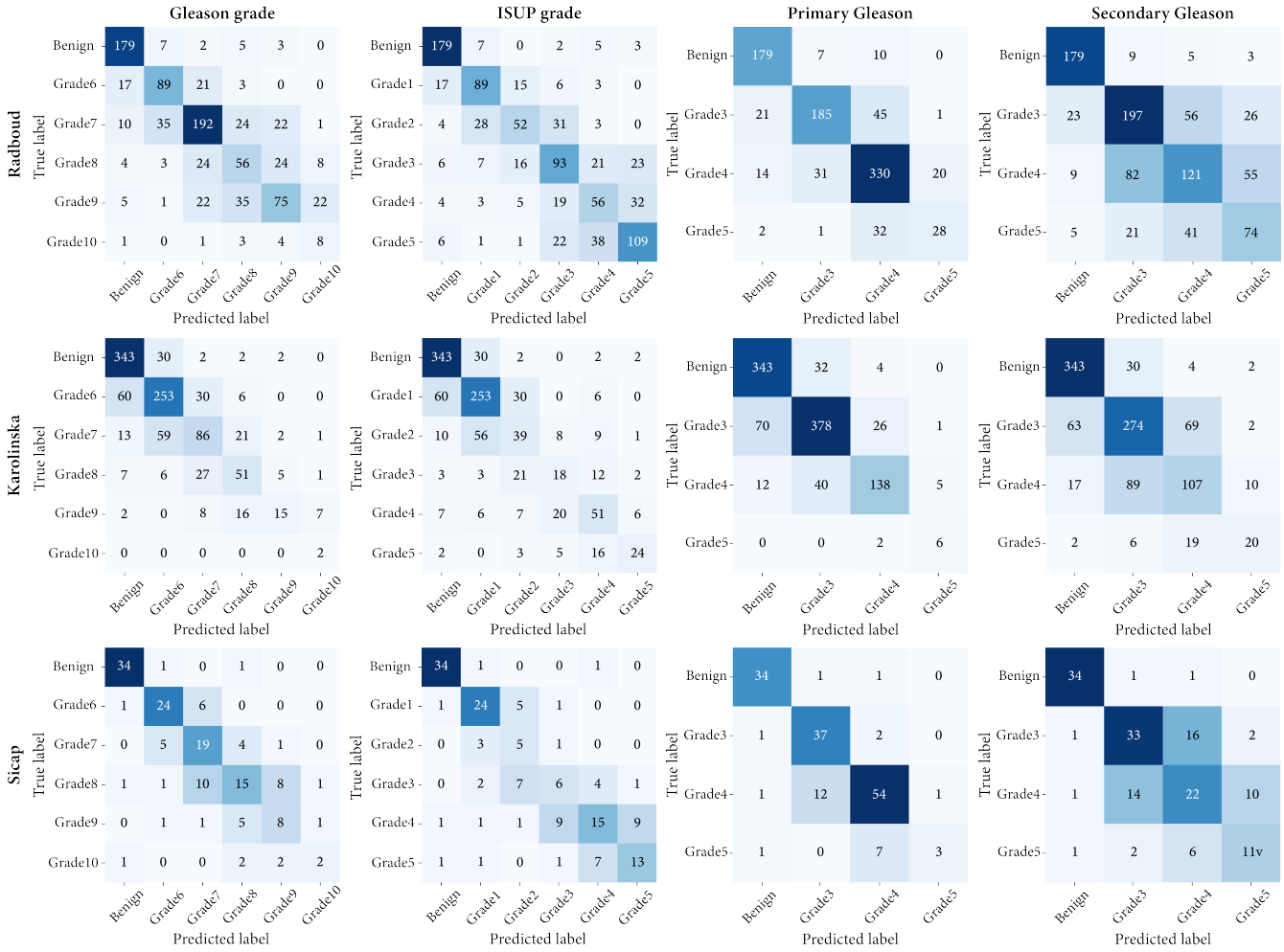

图3：采用 Radboud 和 Karolinska 训练数据集训练的最佳 WHOLESIGHT 模型，在 Radboud、Karolinska 和 Sicap 数据集上的格里森分级、ISUP分级、主要和次要格里森分类混淆矩阵。

还展示了分割的泛化能力评估。WHOLESIGHT 的表现始终优于 WHOLESIGHT（图、GRAPHGRAD-CAM）。值得注意的是，Radboud 和 Sicap 数据集上的 Dice分数较表II 和 I中的分割结果分别提高了 5.2% 和 4.4%，其中 WHOLESIGHT 的提升尤为明显。这可能归因于使用了更多训练集全切片图像，这表明弱监督语义分割 可通过利用更多弱监督得到改进。

不确定性分析：图 4 展示了 WHOLESIGHT、WHOLESIGHT（图、GRAD-CAM）以及 WHOLESIGHT（多重、NC）在 Radboud 和 Karolinska 数据集上的分类不确定性分析，采用 负对数似然 和 Brier分数进行评估。WHOLESIGHT（多重、NC）在所有数据集的主要、次要和 Gleason 分级（ P+S ）分类中，其 负对数似然 均显著低于 WHOLESIGHT。值得注意的是，在预测次要 Gleason 模式时，负对数似然和布里尔分数始终高于预测主要模式时的结果。这与次要模式识别更具挑战性且存在更高歧义这一事实相一致。

模型校准分析：不确定性估计良好的模型应经过良好校准，即，模型的置信度应接近模型的性能。图4展示了 Karolinska 和 Radboud 数据集上主要分类头的可靠性图。WHOLESIGHT 的校准表现始终优于 WHOLESIGHT（图，GRAD-CAM），且与 WHOLESIGHT（多重，NC）的校准表现相近。ECE 指标也从定量角度支持了这一观察。然而，我们发现所有模型仍然过度自信，因为各置信度区间内的模型准确率均低于预期校准水平（蓝色所示）。

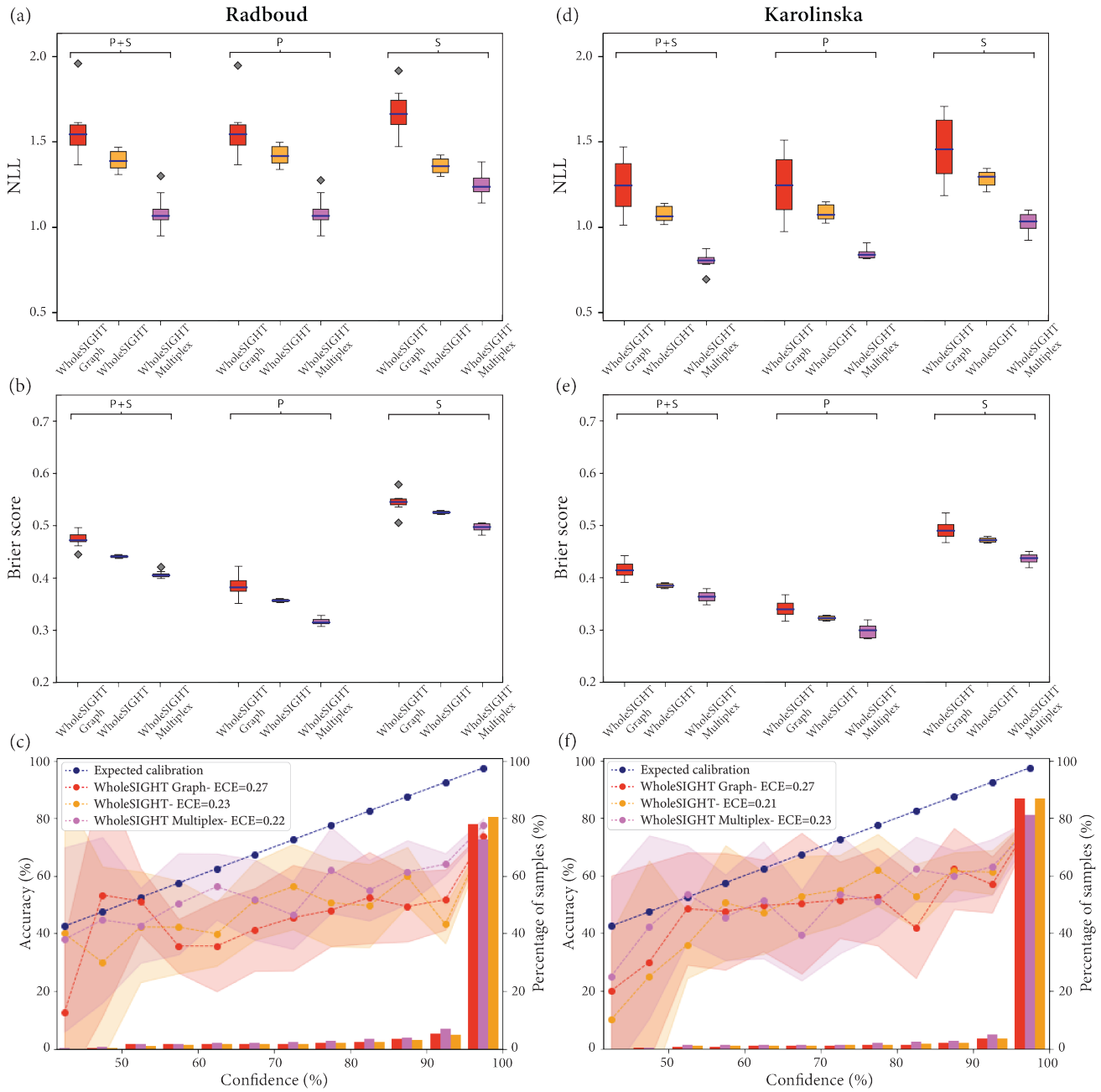

图 4：WHOLESIGHT、WHOLESIGHT（图，梯度加权类激活映射）和 WHOLESIGHT（多重染色，NC）模型在 Radboud（a、b、c）和 Karolinska（d、e、f）数据集上的不确定性与模型校准分析。(a、d) 和 (b、e) 分别展示负对数似然（越低越好）和布里尔分数（越低越好）。(c、f) 展示主要格里森分类头的可靠性图。预期校准（蓝色）表示模型经过完美校准。WHOLESIGHT、WHOLESIGHT（图，梯度加权类激活映射）和 WHOLESIGHT（多重染色，NC）的校准结果分别以橙色、红色和紫色表示，同时标示每个分箱中的样本数量（%）。

F. 定性分析

我们通过以下方式对 WHOLESIGHT 的结果进行定性分析：(1) 在全切片图像上可视化叠加的分割掩膜；(2) 分析t-分布随机邻居嵌入（t-SNE） [88] 节点嵌入，以及 (3) 将分割输出与病理学推理相关联。

分割掩膜可视化：图 5 展示了在 Sicap 数据集上使用 WHOLESIGHT 和 WHOLESIGHT（多重染色，NC）获得的分割预测。我们可以

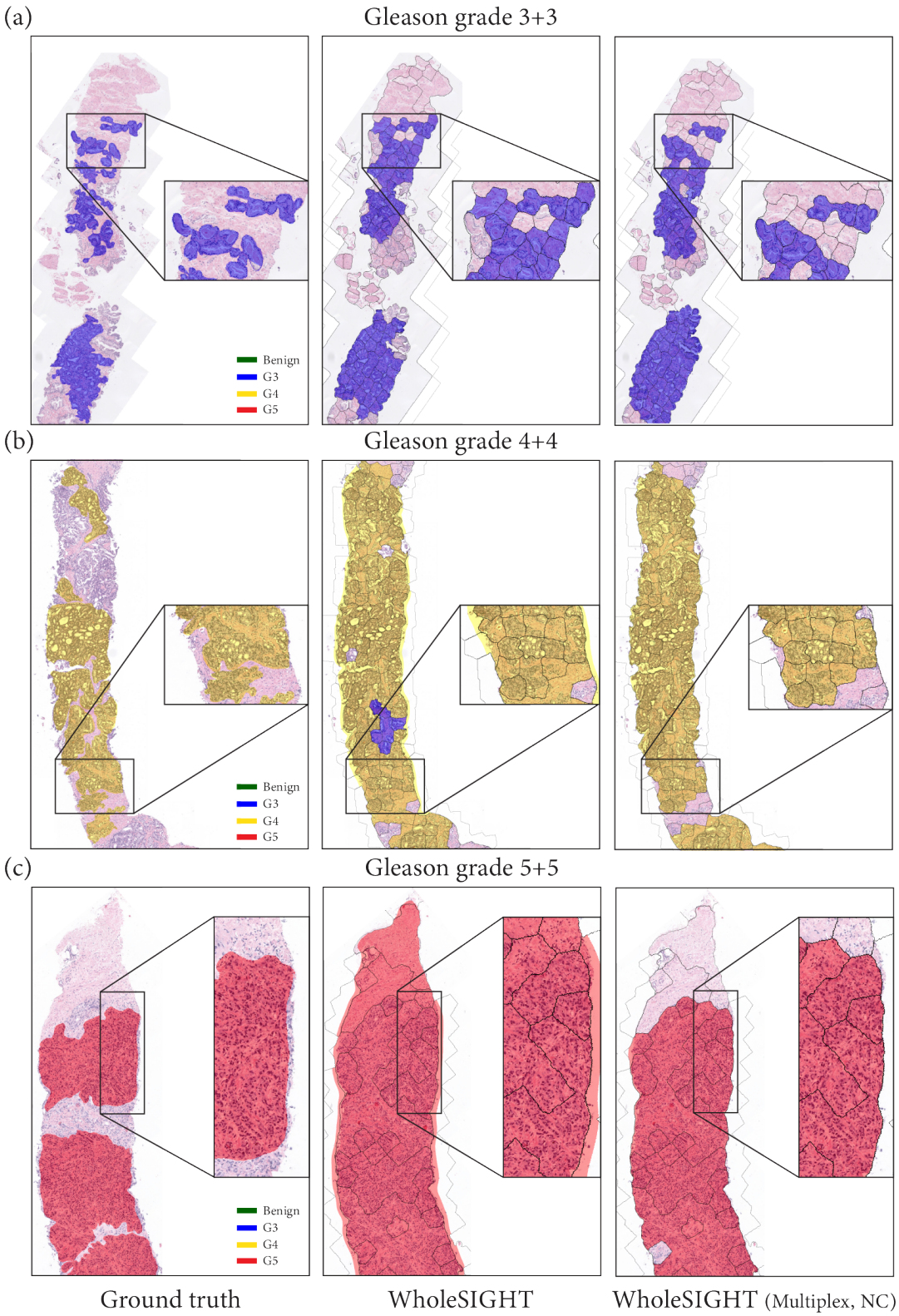

图5：Sicap 数据集中的分割图示例。左侧显示真实标注，中间显示 WholeSIGHT 的预测结果，右侧显示 WholeSIGHT（多重，NC）的预测结果。组织区域（即 *i.e.*，组织图 节点）以黑色覆盖层表示。(a、b、c) 分别展示腺体分级(3+3)、腺体分级(4+4)和腺体分级(5+5)的样本。为便于可视化，分割图中未突出显示良性区域。

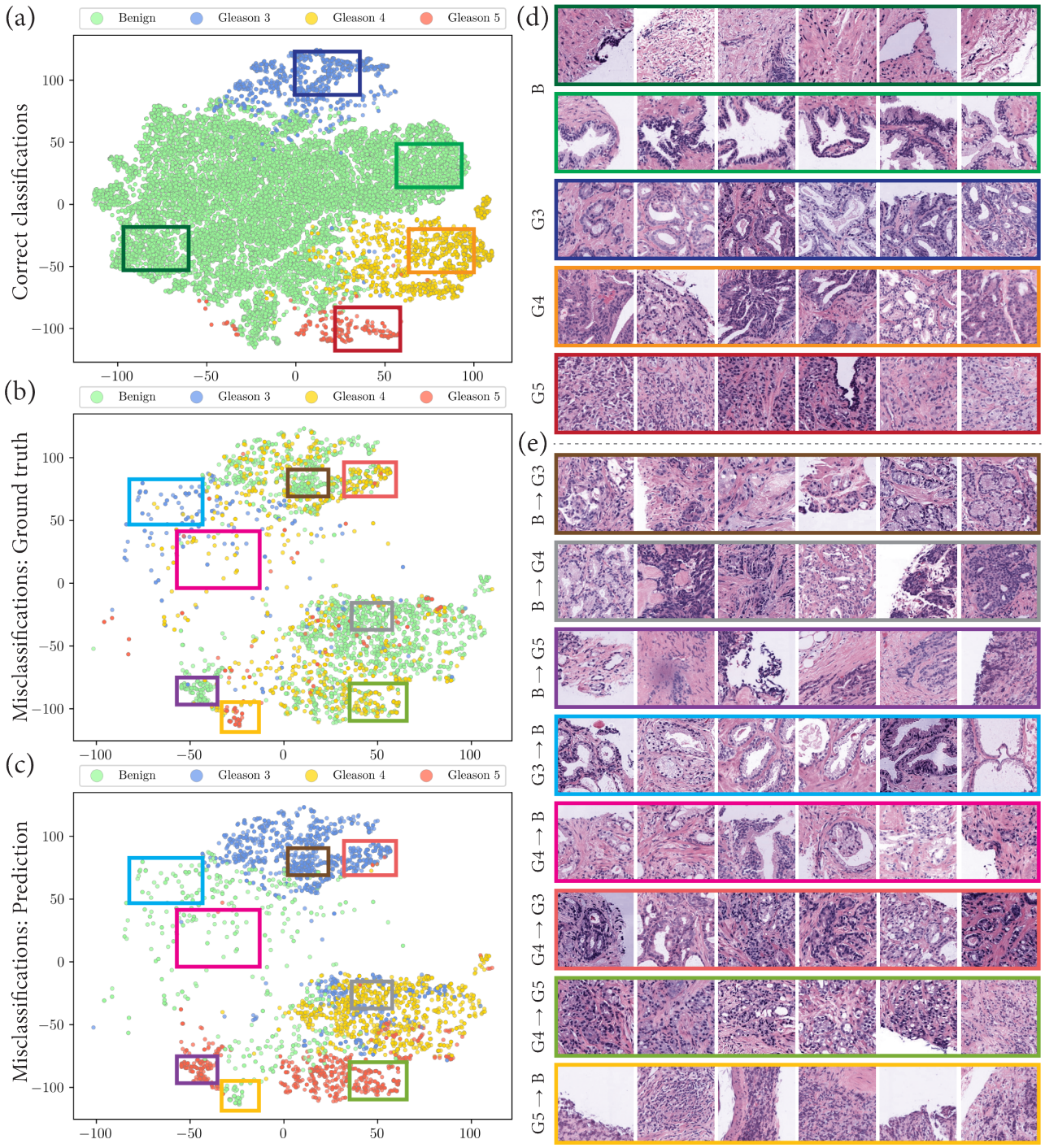

图6：Sicap 数据集的二维 t-SNE 特征空间中，组织图节点嵌入的 t-SNE 可视化结果，以及来自多个区域的示例图像块。(a) 正确分类节点的 t-SNE 可视化结果。(b) 和 (c) 展示了错误分类节点的 t-SNE 可视化结果，其中 (b) 和 (c) 分别突出显示真实标注和预测节点标签。(d) 和 (e) 展示了从 t-SNE 嵌入空间不同区域选取的节点质心周围裁剪出的、大小为 224×224、放大倍率为 $10\times$ 的正方形图像块。(d) 和 (e) 分别突出显示正确分类和错误分类的节点图像块。(e) 中图像块的标签格式为 $Y \rightarrow Y$，其中 Y 和 Y 分别表示真实标注和预测类别标签。(d) 和 (e) 中图像块周围的彩色矩形分别对应于 (a)、(b) 和 (c) 中颜色相同的矩形。

可以看到，WHOLESIGHT能够准确勾勒出全切片图像中的癌变区域。通过放大观察不同区域，可以得出结论：组织图组织图的组织区域（即，组织图的节点组织图，在图5中以黑色轮廓标出）编码了有意义的均质组织单元。这证实了利用组织图表示将组织区域分割为Gleason模式的合理性。我们还注意到，在少数情况下，WHOLESIGHT会将邻近癌变模式的良性区域预测为癌变区域。例如，图5(c)中主要由间质构成的良性区域被预测为G5。我们认为，这些假阳性检测不会妨碍该方法的适用性，因为邻近的癌变区域均能被正确检测。在其他少数情况下，WHOLESIGHT能够正确检测出真实标注中遗漏的癌变区域。例如，图5(b)中全切片图像上部遗漏的G4区域被正确识别。

将WHOLESIGHT与WHOLESIGHT（多重染色，NC）进行比较，我们发现后者消除了若干假阳性，例如图 即如图 5(a)中所示，因此分割结果更加准确。然而，WHOLESIGHT（多重染色，NC）所取得的改进，是以使用像素级标注进行训练为代价的，而这类标注在现实应用中几乎无法获取。因此，WHOLESIGHT似乎是在分割性能与标注需求之间颇具吸引力的折中方案。

特征空间可视化：*t-SNE*： Sicap 的 t-SNE 可视化展示了学习到的组织级嵌入，如图6 所示。 t-SNE将图神经网络 节点嵌入投影到二维特征空间，从而能够分析节点嵌入与 Gleason模式分布之间的联系。

图 6(a) 展示了 t-SNE 特征空间中被正确分类的节点，显示出每种 Gleason模式各自清晰的聚类。良性节点形成的大型聚类表明良性组织具有多样性。图6(d) 展示了每种 Gleason模式聚类中的若干图像块。我们可以观察到，从良性组织到 Gleason 5级，图像块中的细胞核分化程度逐渐降低。此外，图6(b) 和 (c) 展示了错误分类节点的 t-SNE特征空间。具体而言，图6(b) 展示真实节点标签，图 6(c) 展示预测节点标签。我们进一步选取不同的嵌入位置，并用不同的彩色矩形标出，再将其与相应的图像块联系起来，以说明类别间歧义，如图 6(e) 所示。例如，图6(e) 第一行展示了良性图像块被预测为 G3 的情况。我们可以将这些图像块与图6(d) 第三行中的 G3 图像块进行视觉比较。图6(e) 还展示了其他Gleason模式之间的类似歧义。

通过预测分割结果解读模型输出：预测分割结果提供了易于人理解的可解释性图谱。对于研究人员而言，这些分割结果有助于：(1) 识别导致全切片图像分类的形态学模式；(2) 通过检查像素级预测分析失败案例；并最终 (3) 更好地理解模型行为，以助力生物标志物发现。对于病理学家而言，这些结果有助于：(1) 将预测的全切片图像级Gleason评分与突出显示的像素级 Gleason模式联系起来；(2) 确认所识别癌变区域的形态学符合既定诊断标准。

此外，从开发AI辅助的人在回路工具的角度来看，能够同时分类 和 分割 全切片图像 的格里森分级系统更符合最新的病理学标准。事实上，格里森分级系统近期的修订 [83]强调了报告各级别所占百分比的重要性，以便更好地对患者进行分层并选择治疗方案 [89]–[92]。通过统计属于各模式的像素数量，可以直接从预测的分割图中得出这些百分比。显然，单纯的全切片图像 分类系统无法提供此类信息。报告各级别的百分比对于模糊和临界病例尤为重要。例如，考虑两位Gleason评分 3+4 相同的患者。当模式4所占比例较低时，例如，占10%，该病例可视为中危癌症，此时主动监测即可 [93]。然而，较高比例的次要模式可能需要采取特定治疗。报告各级别的百分比有助于我们轻松区分这两种情况。

同样，考虑一个 Gleason评分 4+3 的病例，其次要格里森模式占比较小，例如，主要和次要模式的面积分别为 90% 和 10%。该病例将被评为 4+3，尽管其结果接近4+4，而后者会导致采用不同的治疗方案。明确报告格里森模式百分比，可以避免此类边界情况。

V. 结论

在全视野组织病理图像中准确勾画各种模式通常需要像素级标注，而在实际场景中很难获取。不过，对诊断具有参考意义的模式进行语义分割，对于疾病诊断和治疗方案选择至关重要。为此，我们提出了一种新颖的弱监督语义分割方法 WHOLESIGHT，仅利用图像级监督即可分割组织病理图像中相关的目标模式。据我们所知，WHOLESIGHT 是首个能够以端到端方式处理任意形状和大小组织病理图像的弱监督语义分割方法。我们在三个公开可用的前列腺针吸活检数据集上评估了所提出的方法，用于 Gleason 分级分类

和 Gleason 模式分割。通过与组织病理学应用领域的最先进方法进行比较，我们证明了 WHOLESIGHT 在分类和分割方面的优越性。此外，WHOLESIGHT 是一种模块化方法，既可利用图像级监督，也可利用像素级监督，同时执行图像分类和分割任务。尽管我们在苏木精-伊红染色的前列腺癌针吸活检上评估了该方法，但这项技术很容易扩展到其他组织类型、成像技术和图像模态。

参考文献

[1] R. Siegel、K. Miller、H. Fuchs 和 A. Jemal，“癌症统计数据，2022年”，*CA*：临床医师癌症杂志，第 72 卷，第 7–33 页，2022年。

[2] M. Wilson、K. Fleming、M. Kuti、L. Looi、N. Lago 和 K. Ru，“病理学与检验医学服务的可及性：亟待弥补的关键缺口”，《柳叶刀》，第 391 卷，1938，2018 年第 10133 期，1927 年，页码 – 。[3] M. Amin、S. Smith、V. Reuter、J. Epstein、D. Grignon、D. Hansel、O. Lin、J. McKenney、R. Montironi、G. Paner、H. Al-Ahmadie、F. Algaba、S. Ali、I. Alvarado-Cabrero、L. Bubendorf、L. Cheng、J. Cheville、G. Kristiansen、R. Cote、B. Delahunt、J. Eble、E. Genega、C. Gulmann、A. Hartmann、C. Langner、A. Lopez-Beltran、C. Magi-Galluzzi、J. Merce、G. Netto、E. Oliva、P. Rao、J. Ro、J. Srigley、S. Tickoo、T. Tsuzuki、S. Umar、T. van der Kwast、R. Young 和 M. Soloway，“执业病理学家的最新进展：泌尿系统疾病国际咨询—欧洲泌尿外科学会膀胱癌咨询”，《现代病理学》，第 28 卷，第 5 期，页码 612–630，2015 年。[4] P. Tan、I. Ellis、K. Allison、E. Brogi、S. Fox、S. Lakhani、A. Lazar、E. Morris、A. Sahin、R. Salgado、A. Sapino、H. Sasano、S. Schnitt、C. Sotiriou、P. Diest、V. White、M. Lokuhetty 和 I. Cree，“2019 年世界卫生组织乳腺肿瘤分类”，《组织病理学》，第 77 卷，第 2 期。[5] D. Gomes、S. Porto、D. Balabram 和 H. Gobbi，“普通病理学家与乳腺病理学专家在诊断乳腺小叶肿瘤、柱状细胞病变、非典型导管增生及导管原位癌方面的观察者间差异”，《诊断病理学》，第 9 卷，第 1 期，页码 1–9，2014 年。[6] J. Elmore、G. Longton、P. Carney、B. Geller、T. Onega、A. Tosteson、H. Nelson、M. Pepe、K. Allison、S. Schnitt、F. O’Malley 和 D. Weaver，“解读乳腺活检标本的病理学家之间的诊断一致性”，《美国医学会杂志》，第 313 卷，第 11 期，页码 1122–1132，2015 年。[7] S. Graham、Q. Vu、S. Raza、A. Azam、Y. Tsang、J. Kwak 和 N. Rajpoot，“Hover-net：多组织学图像中细胞核的同步分割与分类”，医学图像分析，第 58 卷，文章编号 101563，2019 年。[8] P. Pati、A. Foncubierta-Rodr´ıguez、O. Goksel 和 M. Gabrani，“减少数字病理学中的标注工作量：用于分类任务的协同表征学习框架”，医学图像分析，第 67 卷，文章编号 101859，2021 年。

[9] S. Graham、Q. D. Vu、M. Jahanifar、S. E. A. Raza、F. Minhas、D. Snead 和 N. Rajpoot，“一个模型就够了：多任务学习实现组织学图像分割与分类同步进行”，医学图像分析，第 83 卷，文章编号 102685，2023 年。[10] K. Sirinukunwattana、J. Pluim、H. Chen、X. Qi、P.-A. Heng、Y. Guo、L. Y. Wang、B. Matuszewski、E. Bruni、U. Sanchez、A. B¨ohm、O. Ronneberger、B. Ben Cheikh、D. Racoceanu、P. Kainz、M. Pfeiffer、M. Urschler、D. Snead 和 N. Rajpoot，“结肠组织学图像中的腺体分割：GLAS 挑战赛”，医学图像分析，第 35 卷，页码 489–502，2017。

[11] T. Binder、E. Tantaoui、P. Pati、R. Catena、A. Set-Aghayan、M. Gabrani，“使用深度学习进行多器官腺体分割，”医学前沿，第 6 卷，第 173 页，2019。

[12] B. Ehteshami Bejnordi、M. Veta、P. Johannes van Diest、B. van Ginneken、N. Karssemeijer、G. Litjens、J. A. W. M. van der Laak 和 CAMELYON16 Consortium，“对用于检测乳腺癌女性患者淋巴结转移的深度学习算法进行诊断评估”，《美国医学会杂志》，第 318 卷，第 22 期，第 2199–2210 页，2019。

[13] G. Aresta、T. Araújo、S. Kwok、S. S. Chennamsetty、M. Safuan、V. Kori、P. Sahiner、B. Ehteshami Bejnordi、N. Tariq、M. Amgad、H. Kang、D. A. Woodruff、P. Gao、J. Yanagihara、J. Sun、L. Gong、G. Shen、T. Zhang、K. Sirinukunwattana、S. Raza、J. Kittler、N. M. Rajpoot、D. A. Snead、J. P. Papa、A. M. Campbell、A. Mehanian、V. Ospina、J. Ho、D. L. Rimm、J. Hipp、A. R. Hanna、O. S. Pinto、C. Bui、B. Shana’ah、A. Gallas、G. Braz、L. Teodoro、A. Fick、V. Madugundu、A. Kumar、P. Jaisakthi、J. Cuevas、A. Adams、S. Vandana、R. Agarwal、G. Sharma、S. B. Marami、M. Prastawa、M. Chan、M. Donovan、G. Fernandez、J. Zeineh、M. Kohl、C. Walz、F. Ludwig、S. Braunewell、M. Baust、Q. Vu、M. To、E. Kim、J. Kwak、S. Galal、V. Sanchez-Freire、N. Brancati、M. Frucci、D. Riccio、Y. Wang、L. Sun、K. Ma、J. Fang、I. Kone、L. Boul-mane、A. Campilho、C. Eloy、A. Pol´onia 和 P. Aguiar，“BACH：乳腺癌组织学图像大挑战”，医学图像分析，第 56 卷，第 122–139 页，2019。

[14] M. Barbieri 和 F. Mahmood，“全切片图像上的数据高效弱监督计算病理学”，《自然-生物医学工程》，第 5 卷，第 6 期，第 555–570 页，2021 年。

[15] G. Campanella、M. Hanna、L. Geneslaw、A. Miraflor、V. Silva、K. Busam、E. Brogi、V. Reuter、D. Klimstra 和 T. Fuchs，“使用全切片图像上的弱监督深度学习进行临床级计算病理学”，《自然-医学》，第 25 卷，第 8 期，第 1301–1309 页，2019 年。

[16] L. Chen、M. Hussein 和 K. Platanios，“图像域中弱监督语义分割的综合分析，”*International Journal of Computer Vision*，第129卷，第2期，第361–384页，2021年。

[17] J. Xie、R. Liu、J. Luttrell IV 和 C. Zhang，“基于深度学习的乳腺癌组织病理图像分析”，第10卷，第80页，2019年。

[18] Y. Xu、Z. Jia、L.-B. Wang、Y. Ai、F. Zhang、M. Lai 和 E. Chang，“通过深度卷积激活特征实现大规模组织病理学图像分类、分割和可视化”，《生物信息学》，第 18 卷，第 281 期，2017 年。

[19] Y. Xu、J.-Y. Zhu、E. Chang、Z. Tu，“弱监督组织病理癌症图像分割与分类”，医学图像分析，第18卷，第3期，第591–604页，2014年。

[20] L. Hou, D. Samaras, T. Kurc, Y. Gao, J. Davis, and J. Saltz, “用于全切片组织图像分类的基于图块的卷积神经网络”，发表于*IEEE/CVF*计算机视觉与模式识别会议 *CVPR*，2016年，第2424–2433页。

[21] Z. Jia, X. Huang, E. I.-C. Chang, 和 Y. Xu, “受约束的深度弱监督组织病理图像分割”，*IEEE* 医学影像汇刊，第36卷，第11期，第2376–2388页，2017年。

[22] G. Xu, Z. Song, Z. Sun, C. Ku, Z. Yang, C. Liu, S. Wang, J. Ma, 和 W. Xu, “Camel：一种用于组织病理图像分割的弱监督学习框架”，发表于国际计算机视觉会议 *(ICCV)*，2019 年，第 10 681–10 690 页。

[23] D. Ho, D. Yarlagadda, T. D’Alfonso, M. Hanna, A. Graben- stetter, P. Ntiamoah, E. Brogi, L. Tan 和 T. Fuchs, “用于多类别乳腺癌图像分割的深度多倍率网络”，计算机化医学影像与图形学，第 88 卷，101866 页，2021 年。

[24] J. Zhang, Z. Zhu, K. Yan, Y. Tian, J. Hu, Z. Liu, 和 X. Han, “用于病理图像弱监督分割的联合全卷积网络和图卷积网络”，医学图像分析，第 73 卷，第 102183 页，2021 年。

[25] C. Han, J. Lin, J. Mai, Y. Wang, Q. Zhang, B. Zhao, X. Chen, X. Pan, Z. Shi, Z. Xu, S. Huang, T. Cao, L. Yan, G. Zhao, J. Cheng, F. Han, J. Guan, L. Wang, 和 D. Shen, “多实例学习中全切片图像的可微缩放”，载于欧洲计算机视觉会议 *(ECCV)*，2022年。

[26] L. Chan, M. Hosseini, C. Rowsell, K. Plataniotis, 和 S. Damaskinos, “HistoSegNet：全切片图像中组织类型的语义分割”，载于国际计算机视觉大会（*ICCV*），2019 年，第 10661–10670 页。

[27] J. Silva-Rodríguez, A. Colomer, 和 V. Naranjo, “Weglenet：用于前列腺组织学图像中 Gleason 分级语义分割的弱监督卷积神经网络”，计算机化医学影像与图形学，第 88 卷，页码 101846，2021 年。

[28] J. Ahn, S. Kwak, 利用像素间关系进行实例分割的弱监督学习”，载于*IEEE/CVF*计算机视觉与模式识别会议（*CVPR*），2019 年，第 2204–2213 页。

[29] V. Anklin, P. Pati, G. Jaume, B. Bozorgtabar, A. Foncubierta- Rodriguez, J. Thiran, M. Sibony, M. Gabrani 和 O. Goksel, “利用组织图从不精确且不完整的标签中学习全视野切片分割”，载于医学图像计算与计算机辅助干预（*MICCAI*），2021 年，第 636–646 页。

[30] M. Ilse, J. Tomczak, 和 M. Welling, “基于注意力的深度多实例学习”，载于国际机器学习会议*(ICML)*，2018年，页码 2127–2136。

[31] A. Myronenko、Z. Xu、D. Yang、H. Roth 和 D. Xu，“在全切片成像中考虑基于深度学习的多实例学习的依赖关系”，载于医学图像计算与计算机辅助干预 *(MICCAI)*，2021，页码 329–338。[32] Z. Shao、H. Bian、Y. Chen、Y. Wang、J. Zhang、X. Ji 和 Y. Zhang，“Transmil：基于 Transformer 的相关多实例学习，用于全切片图像分类”，载于神经信息处理系统（*NeurIPS*）进展，2021。[33] Y. Lee、J. Park、S. Oh、K. Shin、J. Sun、M. Jung、C. Lee、H. Kim、J. Chung、K. Moon 和 S. Kwon，“通过图深度学习从肿瘤全切片图像中提取预后相关的上下文组织病理学特征”，《自然-生物医学工程》，页码 1–15，2022年。

[34] J. Lindvall, T. Chew, A. Sahasrabudhe, I. Wen, Y. Hong, J. Wang, Z. Liu, Q. Sun, K. Yoon, S. Wong, A. Laquian, C. Hsieh, D. Gratzinger, M. Ohgami, B. King, A. Rabinovitch, J. Shin, J. Kohlmeyer, M. Ko, 和 A. Natkunam, “Quantitative analysis of patient-specific morphological features of follicular lymphoma using deep learning”, *Modern Pathology*, 第 37 卷，第 100 483 页，2024 年。

[35] K. Thandiackal, B. Chidester, P. Pati, K. Kanev, D. Demeter, J. Portenier, A. Gabrani, M. Do, 和 O. Goksel, “Differentiable zooming for multiple-instance learning on whole-slide images”, 载于*IEEE/CVF* 计算机视觉与模式识别会议 *(CVPR)*，2022年，第 2384–2394 页。

[36] F. Williamson, M. Gabrani, 和 O. Goksel, “Discrete fourier descriptors for preserving anisotropic contour information”, *Pattern Recognition*, 第 122 卷，第 108 310 页，2022 年。

[37] H. Lee 和 W. Jeong, “Scribble2label：通过一致性约束自生成伪标签的涂鸦监督细胞分割”，载于医学图像计算与计算机辅助手术干预（*MICCAI*），2020年，第14–23页。

[38] H. Qu, P. Wu, Q. Huang, J. Yi, Z. Yan, K. Li, G. M. Riedlinger, S. De, S. Zhang, 和 D. N. Metaxas, “使用部分点标注对组织病理图像进行弱监督深度细胞核分割”，*IEEE* 医学影像汇刊，第39卷，第11期，第3655–3666页，2020年。

[39] T.-A. Yen, H.-C. Hsu, P. Pati, M. Gabrani, A. Foncubierta- Rodríguez, and P.-C. Chung, “Ninepins：基于点标注的细胞核实例分割，” *arXiv* 预印本 *arXiv:2006.13556*，2020年。

[40] J.-M. Nagtegaal、J. van der Laak、F. Ciompi, 《从稀疏标注数据中学习组织病理图像的语义分割》，载于医学影像与深度学习（*MIDL*），2018年。

[41] O. Ronneberger, P. Fischer, 和 T. Brox, “U-net：用于生物医学图像分割的卷积网络”，载于医学图像计算与计算机辅助干预——*MICCAI 2015*，N. Navab、J. Hornegger、W. M. Wells 和 A. F. Frangi 编。Cham：施普林格国际出版，2015年，第234–241页。

[42] M. Aubreville，“量化有丝分裂检测中的扫描仪引起的域差距”，载于医学影像与深度学习（*MIDL*），2021。

[43] M . Bertram, J. Wooley, X. Guan, C. Schmitt, and N. Thomas，“一种用于组织学切片定量分析的标准化方法”，载于*IEEE*国际生物医学成像研讨会（*ISBI*），2009，第1107–1110页。

[44] A . Vahadane, T. Peng, S. Albarqouni, M. Baust, K. Steiger, A. Schlitter, A. Sethi, I. Esposito, 和 N. Navab，《组织学图像的结构保留型颜色归一化与稀疏染色分离》，*IEEE* 医学成像汇刊，第 35 卷，第 8 期，第 1962–1971 页，2016 年。

[45] M . Anand, P. Pati, H. Ruschoff, P. Wild, M. Gabrani, and H. Pozidis，“用于组织病理学全切片图像染色归一化的快速、可扩展流程”，载于欧洲计算机视觉会议（*ECCV*）研讨会，2018年。

[46] J . Ren, I. Hacihaliloglu, E. Singer, D. Foran, 和 X. Qi，“用于组织病理学全切片图像分类的无监督域适应”，*Frontiers in Bioengineering and Biotechnology*，第 7 卷，第 102 页，2019。

[47] D . Tellez, M. Balkenhol, N. Karssemeijer, G. Litjens, J. van der Laak, 和 F. Ciompi，“H&E 染色增强提升了卷积网络在组织病理学有丝分裂检测中的泛化能力”，载于 医学影像，第 10581 卷，2018年，第 264–270 页。

[48] K . Faryna, J. van der Laak, 和 G. Litjens，《针对基于 H&E 染色组织病理学定制自动化数据增强》，载于*Medical Imaging with DeepLearning (MIDL)*，2021年。

[49] S . Onder, M. Argin, V. Andreev, 和 K. Kahveci，《用于改进计算病理学中深度卷积神经网络泛化能力的染色不变特征》，*Frontiers in Bioengineering and Biotech- nology*，2019年。

[50] R . Yamashita, J. Long, S. Banda, J. Shen, and D. L. Rubin，《利用医学无关的风格迁移增强，为计算病理学学习领域无关视觉表示》，*IEEE* 医学影像汇刊，第40卷，第12期，第3945–3954页，2021年。

[51] Y . Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, 和 V. Lempitsky，“神经网络的领域对抗训练”，机器学习研究杂志，第 17 卷，2030，2016 1，2096，第 – 页。

[52] M . Aubreville, C. Bertram, R. Klopfleisch, 和 A. Maier，“用于有丝分裂像评估的跨物种、跨组织域适应：从老方法中学新技巧”，医学图像处理

2020，。2020

[53] N . Hashimoto, D. Fukushima, R. Koga, Y. Takagi, K. Ko, K. Kohno, M. Nakaguro, S. Nakamura, H. Hontani 和 I. Takeuchi，“利用未标注组织病理像进行癌症亚型分类的多尺度域对抗多实例卷积神经网络”，发表于*IEEE/CVF* 计算机视觉与模式识别会议

计算机视觉与模式识别会议（*CVPR*），2020年，第3852–3861 页。[54] M. E. Tschuchnig、G. J. Oostingh 和 M. Gadermayr，“数字病理学中的生成对抗网络：趋势与未来潜力综述”，*Patterns*，第 1 卷，第 6 期，2020年。[55] C. Guo、G. Pleiss、Y. Sun 和 K. Weinberger，“现代神经网络的校准”，发表于国际机器学习会议（*ICML*），2019年，第1321–1330 页。[56] B. Lakshminarayanan、A. Pritzel 和 C. Blundell，“利用深度集成进行简单且可扩展的预测不确定性估计”，发表于神经信息处理系统进展（*NeurIPS*），2017。

[57] S . Fort, H. Hu, 和 B. Lakshminarayanan，《深度集成：损失景观视角》，发表于神经信息处理系统进展（*NeurIPS*），2019。

[58] W . L. Hamilton，《图表示学习》，人工智能与机器学习综合讲座，第14卷，第3期，第1–159页，2020年。

[59] K . Xu, W. Hu, J. Leskovec, 和 S. Jegelka，“图神经网络有多强大？”载于国际学习表征会议（*ICLR*），2019。

[60] J . Nixon, M. Dusenberry, L. Zhang, G. Jerfel, 和 D. Tran，“测量深度学习中的校准”，载于*CVPR*研讨会，第 2 卷，第 7 期，2019 年。

[61] M . DeGroot 和 S. Fienberg，“预测者的比较与评估”，英国皇家统计学会会刊：*D*辑（统计学家），第 32 卷，第 1-2 期，第 12–22 页，1983 年。

[62] M . Veta, J. Pluim, P. Diest, 和 M. Viergever，《乳腺组织病理图像分析：综述》，*IEEE*生物医学工程学报，第 5 期，第 1400–1411 页，2014年。

[63] D . Tellez, G. Litjens, P. Bándi, W. Bulten, J. Bokhorst, F. Ciompi 和 J. van der Laak，《量化数据增强和染色颜色归一化对计算病理学卷积神经网络的影响》，医学图像分析，第 58 卷，第 101544 页，2019 年。

[64] B . Bejnordi, G. Litjens, M. Hermsen, N. Karssemeijer 和 J. van der Laak，“一种用于在全切片组织病理图像中检测感兴趣区域的多尺度超像素分类方法”，载于医学影像学*2015*：数字病理学，第 9420 卷，2015年，第 99–104 页。

[65] P . Pati, G. Jaume, A. Foncubierta-Rodríguez, F. Feroce, G. Scognamiglio, N. Brancati, M. Fiche, E. Dubruc, D. Riccio, M. Bonito, G. Pietro, G. Botti, J. Thi- ran, M. Frucci, O. Goksel 和 M. Gabrani，《数字病理学中全切片图像的解读》，2022年，美国专利申请 16/953,377。

[66] P . Pati, G. Jaume, A. Foncubierta-Rodríguez, F. Feroce, G. Scognamiglio, N. Brancati, M. Fiche, E. Dubruc, D. Riccio, M. Bonito, G. Pietro, G. Botti, J. Thi- ran, M. Frucci, O. Goksel 和 M. Gabrani，“数字病理学中的层次图表示”，医学图像分析，第 75 卷，102264 页，2022年。

[67] R. Achanta, A. Shaji, K. Smith, Lucchi, P. Fua 和 S. Susstrunk，“与最先进的超像素方法相比的简单线性迭代聚类超像素”，*IEEE* 模式分析与机器智能汇刊，第 34 卷，2282，2012 年 11 月，第 2274– 页。

[68] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov 和 L.-C. Chen，“Mobilenetv2：倒置残差与线性瓶颈”，发表于 *IEEE/CVF Conference on Computer Vision and PatternRecognition(CVPR)*，2018年，第4510–4520页。

[69] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and L. Fei-Fei，《ImageNet：一个大规模分层图像数据库》，发表于 *IEEE/CVF*计算机视觉与模式识别会议（*CVPR*），2009年，第248–255页。

[70] F. Potjer，“区域邻接图与连通形态学算子”，载于数学形态学及其在图像与信号处理中的应用。计算成像与视觉，第5卷，1996年，第111–118页。

[71] K. Xu, C. Li, Y. Tian, T. Sonobe, K. Kawarabayashi, and S. Jegelka，“图表示学习：基于跳跃知识网络的图表示学习”，发表于 国际机器学习会议 *(ICML)*，2018年。

[72] P. Pope, S. Kolouri, M. Rostami, C. Martin, 以及 H. Hoffmann，“图卷积神经网络的可解释性方法”，载于*IEEE/CVF*计算机视觉与模式识别会议 *(CVPR)*，2019，第 10 764–10773 页。

[73] G. Jaume, P. Pati, B. Bozorgtabar, A. Foncubierta-Rodríguez, F. Feroce, A. Anniciello, T. Rau, J. Thiran, M. Gabrani, and O. Goksel，“量化计算病理学中图神经网络解释器”，收录于*IEEE/CVF*计算机视觉与模式识别会议（*CVPR*），2021年，第8106–8116页。

[74] R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, and D. Ba- tra，“梯度加权类激活映射：来自深度网络的可视化解释”，发表于国际计算机视觉大会 *(ICCV)*，2017，第 618–626 页。

[75] J. Silva-Rodríguez, A. Colomer, M. Sales, R. Molina, 以及 V. Naranjo，“深入格里森评分量表：用于前列腺组织学分级和筛状结构检测的自动端到端系统”，《生物医学中的计算机方法与程序》，第 195 卷，第 105637 页，2020年。

[76] W. Bulten, H. Pinckaers, H. Boven, R. Vink, T. Bel, B. Gin- neken, J. van der Laak, C. Hulsbergen-van de Kaa, 和 G. Litjens，“基于活检对前列腺癌进行格里森分级的自动化深度学习系统：一项诊断研究”，《柳叶刀·肿瘤学》，第 21 卷，第 2 期，第 233–241 页，2020年。

[77] W. Bulten, K. Kartasalo, P. Y. Chen, J. Ström, H. Pinckaers, K. Nagpal, Y. Cai, D. Steiner, H. van Boven、R. Vink、C. Hulsbergen-van de Kaa、J. van der Laak、M. Amin、A. Evans、T. van der Kwast、R. Allan、P. Humphrey、H. Grönberg、H. Samaratunga、B. Delahunt、T. Tsuzuki、T. Häkkinen、L. Egevad、M. Demkin、S. Dane、F. Tan、M.Valkonen、G.Corrado、L.Peng、C.Mermel、P.Ruusuvuori、G. Litjens、M. Eklund及PANDA挑战赛联盟，“用于前列腺癌诊断和格里森分级的人工智能：PANDA挑战赛”，《自然-医学》，第28卷，第154–163页，2022年。

[78] P. Ström、K. Kartasalo、H. Olsson、L. Solorzano、B. Delahunt、D. Berney、D. Bostwick、A. Evans、D. Grignon、P. Humphrey、K. Iczkowski、J. Kench、G. Kristiansen、T. Van der Kwast、K. Leite、J. McKenney、J. Oxley、C.-C. Pan、H. Samaratunga及M. Eklund，“利用人工智能进行病理学家级别的前列腺活检样本分级”，《生物信息学》，2019年。[79] A. Paszke、S. Gross、F. Massa、A. Lerer、J. Bradbury、G. Chanan、T. Killeen、Z. Lin、N. Gimelshein、L. Antiga、A. Desmaison、A. Köpf、E. Yang、Z. DeVito、M. Raison、A. Tejani、S. Chilamkurthy、B. Steiner、L. Fang及S. Chin- tala，“PyTorch：一种命令式风格的高性能深度学习库”，载于神经信息处理系统（*NeurIPS*），2019年，第8024–8035页。

[80] M. Wang, L. Yu, D. Zheng, Q. Gan, Y. Gai, Z. Ye, M. Li, J. Zhou, Q. Huang, C. Ma, Z. Huang, Q. Guo, H. Zhang, H. Lin, J. Zhao, J. Li, A. Smola, and Z. Zhang，“深度图书馆：面向图上的高效且可扩展的深度学习”，发表于国际学习表征会议 *(ICLR) workshop*，2019。

[81] G. Jaume, P. Pati, V. Anklin, A. Foncubierta, and M. Gabrani，“Histocartography：用于数字病理学图分析的工具包”，载于*MICCAI*计算病理学研讨会（*MICCAI-W*），2021，pp.117–128。

[82] J. Epstein、W. Allsbrook、M. Amin、L. Egevad和ISUP分级委员会，“国际泌尿病理学会（ISUP）关于前列腺癌格里森分级的2005年共识会议”，《美国外科病理学杂志》，第29卷，第9期，pp. 1228–1242，2005。

[83] J. Epstein、L. Egevad、M. Amin、B. Delahunt、J. Srigley、P. Humphrey和分级委员会，“2014年国际泌尿病理学会（ISUP）关于前列腺癌格里森分级的共识会议：分级模式的定义及新分级系统的提案”，《美国外科病理学杂志》，第40卷，第2期，第244–252页，2016年。

[84] A. Gomariz, T. Portenier, C. Nombela-Arrieta, 和 O. Goksel，《定量显微镜中的概率空间分析：通过密度图的深度贝叶斯回归实现不确定性感知细胞检测》，《科学进展》，第 8 卷，第 5 期，2021年。

[85] G. Jacobs, “Zhang 等人，“一种结构化网络的核的度量嵌入可训练校准度量”，载于国际机器学习会议 *(ICML)*，2018年。

[86] T. A. Ozkan, A. T. Eruyar, O. O. Cebeci, O. Memik, L. Oz- can, and I. Kuskonmaz，《前列腺癌格里森组织学分级中的观察者间变异》，《斯堪的纳维亚泌尿学杂志》，第50卷，第6期，第420–424页，2016年。

[87] E. N.Salmo，《前列腺癌活检格里森分级中观察者间变异的审计：英格兰西北部中央病理复核的经验》，*Integr Cancer Sci Ther*，第2卷，第2期，第104–106页，2015年。

[88] L. 范德马滕和 G. Hinton，“使用 t-SNE 可视化数据”，机器学习研究期刊，第 9 卷，第 11 期，2008 年。

[89] L. Cheng、D. Davidson、H. Lin 和 M. Koch，“Gleason 4级模式和 5 级模式的占比可预测根治性前列腺切除术后的生存情况”，《癌症》，第 110 卷，第 9 期，第 1967–1972 页，2007 年。

[90] C. Hyang M、Kong M、Zhou A、Rosenberg S、Tanaja、J. Melamed 和 F. Deng，“计数活检中 Gleason评分为 3+4=7 、仅含少量 Gleason 4级模式的前列腺癌，在根治性前列腺切除术标本中与低危肿瘤相关”，《美国外科病理学杂志》，第 38 卷第 8 期，第 1096–1101 页，2014年。

[91] B. Chou S、Pearce B、Anderson A、Skalbaa G、Zagaja, S. Eggener 和 G. Paner，“根治性前列腺切除术中当代 Gleason 4级模式前列腺癌的百分比及组织结构类型的预后意义”，《美国外科病理学杂志》，第 40 卷，第 10 期，第 1400–1406 页，2016 年。

[92] M. Shiyamoto H、Miyamoto，“Gleason 4级模式百分比在中危前列腺癌患者预后分层中的作用”，《转化男科学与泌尿学》，第 7 卷，第 4 期，2018年。

[93] M. Amin P、Lin L、Gore J、Srigley H、Samaratunga, L. Egevad, M. Rubin, J. Nacey, H. Carter, L. Klotz, H. San- dler, A. Zietman, S. Holden, R. Montironi, P. Humphrey, A. Evans, J. Epstein, B. Delahunt, J. McKenney, D. Berney, T. Wheeler, A. Chinnaiyan, L. True, B. Knudsen, 和 M. Hammond，“病理学家在确定前列腺癌患者主动监测这一管理方案的适用资格方面发挥着关键作用：共识声明及建议，获美国病理学家学会、国际泌尿病理学会、解剖与外科病理学主任协会、新西兰病理学家学会和前列腺癌基金会支持”，《病理学与实验室医学档案》，第138卷，第10期，第1387–1405页，2014年。