公共CT数据上的胰腺病灶分割会在哪些地方失败？对 PanTS 上导管引导的流水线进行按衰减分层分析

Kang-Won Byun（独立研究者），bgw2001@gmail.com

摘要

背景。病灶分割是大多数用于胰腺癌检测的计算机断层扫描流水线的基础环节，但病灶重叠评分仍然偏低。我们没有提出新的架构，而是报告端到端流水线在大型公共数据集上的失效之处。

方法。我们在 PanTS 上构建了一个三阶段流水线：胰腺分割、主胰管（MPD）和胆总管（CBD）的导管图，以及一个接收含或不含预测胰管图的计算机断层扫描图像的病灶模型。两个病灶模型均以相同方式在 4007 张扫描图像（其中 417 张有病灶）上训练，在 150 张验证集扫描图像上进行选择，并根据测试运行前确定的比较规则，在 737 张留出测试扫描图像（其中 127 张有病灶）上评估一次。我们利用 168 张有病灶和 1433 张无病灶的训练扫描图像的折外预测来定位失效：预言式移除假阳性连通区域、病灶覆盖率、病灶相对于正常胰腺的衰减、候选筛选，以及作为病灶存在情况单独征象的主胰管和胆总管直径。

结果。 在测试集上，使用导管图时病灶Dice系数为 0.213（95%置信区间 0.173 至 0.259），仅CT时为 0.209（95%置信区间 0.169 至 0.252）（配对差异 +0.004（95%置信区间 -0.011 至 +0.019））。导管图提高了被预测结果触及的病灶比例（+0.079（95%置信区间 +0.024 至 +0.134）），但也扩大了无病灶扫描中的预测区域。在 100.0% 的无病灶测试扫描中，均预测出了病灶样区域。折外分析显示，仅保留触及病灶的预测组件，可使Dice系数从 0.285 提高至 0.309。触及病灶的组件覆盖了病灶中位数 56.4% 的区域，其中位数 29.0% 的体积为病灶。仅凭强度即可与正常胰腺区分的明显低衰减病灶，其覆盖率最低（中位数 16.6% ）。基于候选裁剪区域的三维卷积神经网络，作为候选过滤器的表现不如连通分量体积。仅根据预测图，主胰管直径可区分病灶与无病灶测试扫描，曲线下面积为 0.794（95%置信区间 0.747 至 0.839）；胆总管直径的区分能力较弱（0.603（95%置信区间 0.549 至 0.658）），将其纳入也无助于改善结果（差异 -0.000（95%置信区间 -0.009 至 +0.010））。主胰管直径至少为 5 mm 时，检验敏感度为 44.1%，特异度为 95.9%。

结论。 在这一情境下，较低的病灶Dice系数主要源于正确定位预测的范围以及无病灶扫描中的预测，而不是病灶扫描中额外的连通分量。折外结果显示，约三分之一的病灶与正常胰腺具有相同的衰减，这限制了任何基于强度的边界所能达到的效果。训练、推理和评估代码已公开。

1. 引言

胰腺导管腺癌（PDAC）通常在晚期才被发现，而增强CT是主要的影像学检查。许多自动检测系统首先对胰腺和候选病灶进行分割，然后对每个候选项进行分类；例如，ePAI会分割胰腺及其周围解剖结构，定位所有潜在病灶，并将每个病灶分类为胰腺导管腺癌、非胰腺导管腺癌或正常[1]。继发征象（如主胰管扩张）已知十分重要：等衰减肿瘤“只有在出现继发影像学表现（例如胰管扩张）时才能被识别”[2]，，而导管和胆总管掩膜已被用作额外的网络输入[3]。

在大型公共数据集上，病灶分割质量的特征仍未得到充分刻画；较低的重叠评分通常被归因于小病灶或等衰减病灶，却缺乏对误差来源的细分。如今，包含胰腺、主胰管、胆总管和病灶体素标注的大型公共数据集，使得进行此类细分成为可能。

本报告分为四个部分：1. 在 PanTS 上构建端到端流水线（胰腺 → 主胰管/胆总管图 → 病灶），其中病灶模型以预测的导管图为输入，并设置一个采用相同训练方式的仅CT对照，在预先指定的规则下于一个未公开的

测试集上进行一次比较；2. 将病灶 Dice 损失分解为假阳性连通区域、漏检病灶以及定位正确的预测范围；3. 按病灶相对于正常胰腺的衰减和病灶大小进行分层分析；4. 关于候选筛选的阴性结果（对裁剪区域使用三维卷积神经网络与使用连通分量体积进行比较），以及对主胰管和胆总管直径作为病灶存在情况单独征象的分析。

2. 相关工作

ePAI 采用由解剖结构分割、病灶定位和病灶分类组成的三级级联流程 [1]。Alves 等人使用 nnU-Net 对胰腺及其周围解剖结构（包括胰管）进行分割，并报告了一个等衰减肿瘤位于“沿扩张的胰管末端延伸处”的病例 [2]。Viviers 等人将“CT 扫描、胰腺和两条导管按通道堆叠”作为肿瘤分割和检测的输入 [3]。直接和间接影像学征象 [4] 以及主胰管扩张 [5] 的检测也已有报道；对于后两项，我们仅查阅了摘要。因此，我们的流程并不声称采用了新的架构。其使用预测的胰管图遵循了 [3]。

3. 数据

PanTS。 我们使用了 PanTS CT 数据集，其中包含胰腺、主胰管、胆总管以及一个“胰腺病灶”类别的体素标注 [6]。PanTS 涵盖胰腺导管腺癌、胰腺神经内分泌肿瘤、胰腺囊性肿瘤和囊性非肿瘤性病变 [6]，，且病灶标注不包含病灶类型，因此下文中的“病灶”指任何已标注的胰腺病灶。与 PANORAMA 数据集重复出现的扫描图像（经像素验证的重复项）已在数据集划分前移除。

数据集划分。训练集包含4007份扫描图像（其中417份含有病灶），验证集包含150份（16份），测试集包含737份（127份）。测试集为PanTS ImageTe分区；训练和验证扫描图像取自ImageTr。PanTS未提供患者标识符，因此按扫描图像进行数据集划分；无法排除同一患者的重复扫描图像出现在不同数据集中。此次研究之前，测试集曾被使用过一次，用于比较胰管分割模型；此前未用于任何病灶模型决策。

4. 方法

4.1 流程

第*1*阶段，胰腺。在训练扫描图像上训练的3D SegResNet [7] （MONAI实现 [8]）在完整的计算机断层扫描上分割胰腺。感兴趣区域（ROI）为预测胰腺加上30 mm边缘，并重采样至0.8 × 0.8 × 1.0 mm。

第*2*阶段，导管图。第二个SegResNet使用带膨胀主胰管的训练扫描图像进行训练，在感兴趣区域内预测主胰管和胆总管；导管图的阈值设为0.5。

第*3*阶段，病灶。两个病灶模型采用相同的 SegResNet 配置（16 个初始滤波器，编码器块 1-2-2-4），预测两个 sigmoid 输出，分别为胰腺（包括病灶）和病灶：

- 仅*CT*（对照）：输入为计算机断层扫描，强度裁剪至 −175...250 亨氏单位并缩放至
- [0, 1]；加入导管图：输入为计算机断层扫描及第2阶段预测的两张二值图。

两者的训练过程完全相同：160 × 160 × 128 按照标签类别比例采样图像块（背景 1、胰腺 1、MPD 1、胆总管 0.5、病灶 1.5），ROI边缘抖动为 15 mm，Dice + 交叉熵损失与批次Dice，采用 AdamW（学习率 2×10^{-4}），批大小为 2，每个训练轮次进行 250 次迭代。原计划训练 300 个训练轮次的余弦调度在两个模型均训练 100 个训练轮次后停止，并保留验证病灶Dice系数最佳的检查点（两者均为第 95 个训练轮次）。在 16 个病灶扫描图像上，加入导管图时的验证病灶Dice系数为 0.276，对照组为 0.271。推理采用滑动窗口；当 sigmoid 输出超过 0.5 时，将体素判定为病灶，预测组件采用 26连通。

图1展示了一个测试扫描图像上的三个阶段。

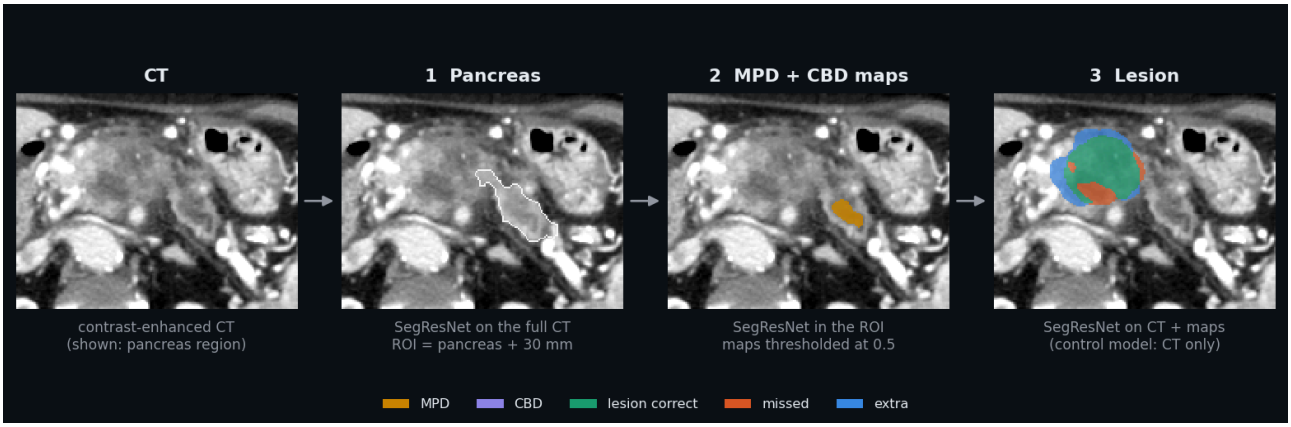


图1。 一个测试扫描（图3中的病例A）上的流程：计算机断层扫描（显示胰腺区域）、第一阶段胰腺、第二阶段主胰管和胆总管图，以及使用导管图的模型的第三阶段病灶预测与参考标注的对比（绿色：正确，橙色：漏检，蓝色：多余预测）。

4.2 用于失败分析的折外预测

训练扫描图像按病灶存在情况分层，划分为五折。训练了两个折模型（导管图模型，采用余弦调度，在100个训练轮数内完成；其余设置同上），并通过完整流水线，对第0折和第1折中的每张扫描图像进行预测，且所用模型均未见过相应扫描图像：168张病灶扫描图像和1433张无病灶扫描图像。由于时间原因，其余三折未进行训练。第1阶段和第2阶段模型均在训练扫描图像上训练过，因此它们在这些扫描图像上的输出属于样本内结果。

4.3 失败分析（折外）

- 神谕组件移除：删除所有未与参考病灶接触的预测组件后得到的病灶Dice系数，即任何组件级筛选器所能达到的上限。覆盖率和病灶占比：对于与病灶接触的组件，其覆盖的病灶比例，以及组件体
- 积中属于病灶的比例。过度移除上限：若保留与病灶接触的组件中仅与病灶重叠的部分，所得的病灶Dice系数。概率阈值：阈值为0.5至0.99时的病灶Dice系数（第0折）。衰减分层：对于每个病灶，随
- 机抽取一个病灶体素，其强度低于随机抽取的正常胰腺体素的概率（体素 AUC）。正常胰腺 = 为经
- 一个体素腐蚀的预测胰腺，减去经两个体素膨胀的病灶，再减去经一个体素膨胀的参考主胰管/胆总管。
- 病灶标记为明显较亮（AUC ≤ 0.4）、相近（0.4–0.6）、较暗（0.6–0.8）或明显较暗（> 0.8）。在与病灶接触的组件内也计算相同的 AUC，比较病灶部分与其余部分（两部分均为 ≥ 20 个体素）。大小分层：根据折外病灶扫描图像的病灶体积分为三分位组（2.15和10.06 ml），并将分界值原样用于测试集。候选筛选：依据以下指标为每个预测组件评分：(i) 体积；(ii) 基于30项组件测量指标（大小、概率、形状、位置、相对于胰腺的强度、与导管图的距离）的梯度提升；(iii) 3D ResNet-10（
- MONAI 实现 [8]），从零开始训练，输入为组件边界框加10 mm范围的 CT裁剪区域，并重采样至
- 643个体素。分类器在第0折和第1折之间进行交叉拟合。设置工作点，使病灶扫描检出率达到80%。导管和胆总管直径（主要测试运行后追加）：预测的主胰管图和预测的胆总管图的最大局部直径（距离变换的两倍），分别作为病灶存在与否的扫描级评分（不区分病灶类型）。使用逻辑回归（输入标准化、默认设置、不调参）组合这两个直径：模型在所有折外训练扫描图像上拟合一次，并应用于测试集一次；是否加入胆总管直径有帮助，则依据配对的测试集 AUC 差异，并采用相同的自助法规则
- 进行判断。为检查是否存在捷径，在移除参考病灶（膨胀一个体素）内部的预测图像体素后，重新计算直径。固定截断值和病灶大小分层均在不拟合的情况下报告。

4.4 测试方案与统计

测试评估方案在运行前已确定并提交：两个病灶模型及其检查点、评估脚本和0.5阈值均与验证集所用设置一致；主要比较指标为病灶Dice系数、病变检出率和最大病变概率的扫描级 AUC 的配对差异（导管图 − 对照）。只有当 95% 自助法置信区间（ 2,000 次病例重采样，随机种子为 0 ）不包含 0 时，才将差异判定为改善。测试集上的结果未用于任何选择。测试集上的衰减分层和大小分层仅作描述性分析。第一阶段胰腺预测为空的扫描图像另行报告。

5. 结果

5.1 测试集：导管图与仅CT

表1和图2列出了测试结果。使用导管图时，病灶Dice系数为0.213（95%置信区间0.173至0.259）；使用仅CT时为0.209（95%置信区间0.169至0.252）；配对差异为+0.004（95%置信区间-0.011至+0.019），检出率差异为 +0.079（95%置信区间 +0.024至 +0.134），扫描级 AUC差异为-0.022（95%置信区间-0.064至+0.022）。根据预先设定的规则，导管图未改变病灶Dice系数或扫描级 AUC，但提高了检出率。检出率的提高伴随着无病灶扫描中的更多预测（使用导管图时预测体积中位数为15.716 ml，而使用仅CT时为12.165 ml）。两种模型几乎都在所有无病灶扫描中预测出某处存在病灶（100.0%和99.8%）。使用仅CT时，在95.3%的病灶扫描和98.7%的无病灶扫描中，最大病变概率至少为0.999（使用导管图时分别为97.6%和99.8%），因此它几乎不包含扫描级信息（AUC为0.468和0.446）；预测体积的表现略好（AUC为0.587和0.545）。一例无病灶扫描因第1阶段胰腺预测为空而失败，另一例因没有参考胰腺标注而在预处理中被排除，最终剩下127例病灶扫描和608例无病灶扫描。

表 1。 测试集（评估了737幅扫描图像；其中127幅为病灶扫描图像，608幅为无病灶扫描图像）。

	仅CT	使用导管图
病灶Dice系数，均值（95%置信区间）	0.209（95%置信区间 0.169 至 0.252）	0.213（95%置信区间 0.173 至 0.259）
病灶Dice系数，中位数	0.079	0.084
发现病灶（预测结果与病灶重叠）	66.9%	74.8%
被预测覆盖的病灶，中位数	24.5%	32.4%
存在任何预测的无病灶扫描图像	99.8%	100.0%
每例无病灶扫描的预测体积，中位数（毫升）	12.165	15.716
每例无病灶扫描的连通分量，中位数	2.000	3.000
预测体积/胰腺体积，无病灶，中位数	15.8%	20.8%
扫描级 AUC，最大病变概率	0.468	0.446
扫描级 AUC，预测病灶体积	0.587	0.545

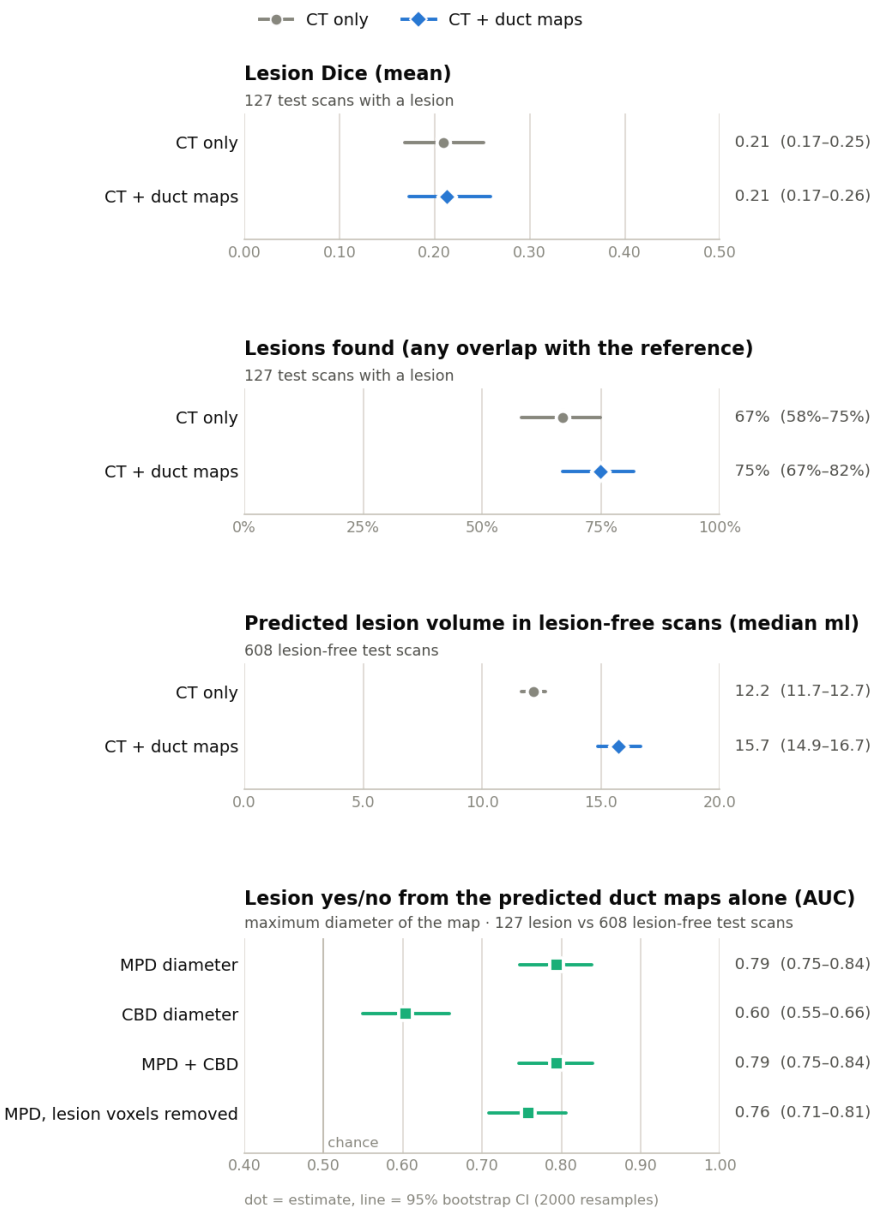


图2。测试集结果（点：估计值；线：95%病例自助法置信区间）。上方三个面板比较两种病灶模型；最下方面板仅显示根据预测图的最大直径得出的扫描级病变存在性（第5.5节）。

图3展示了根据固定规则选出的四个测试扫描。

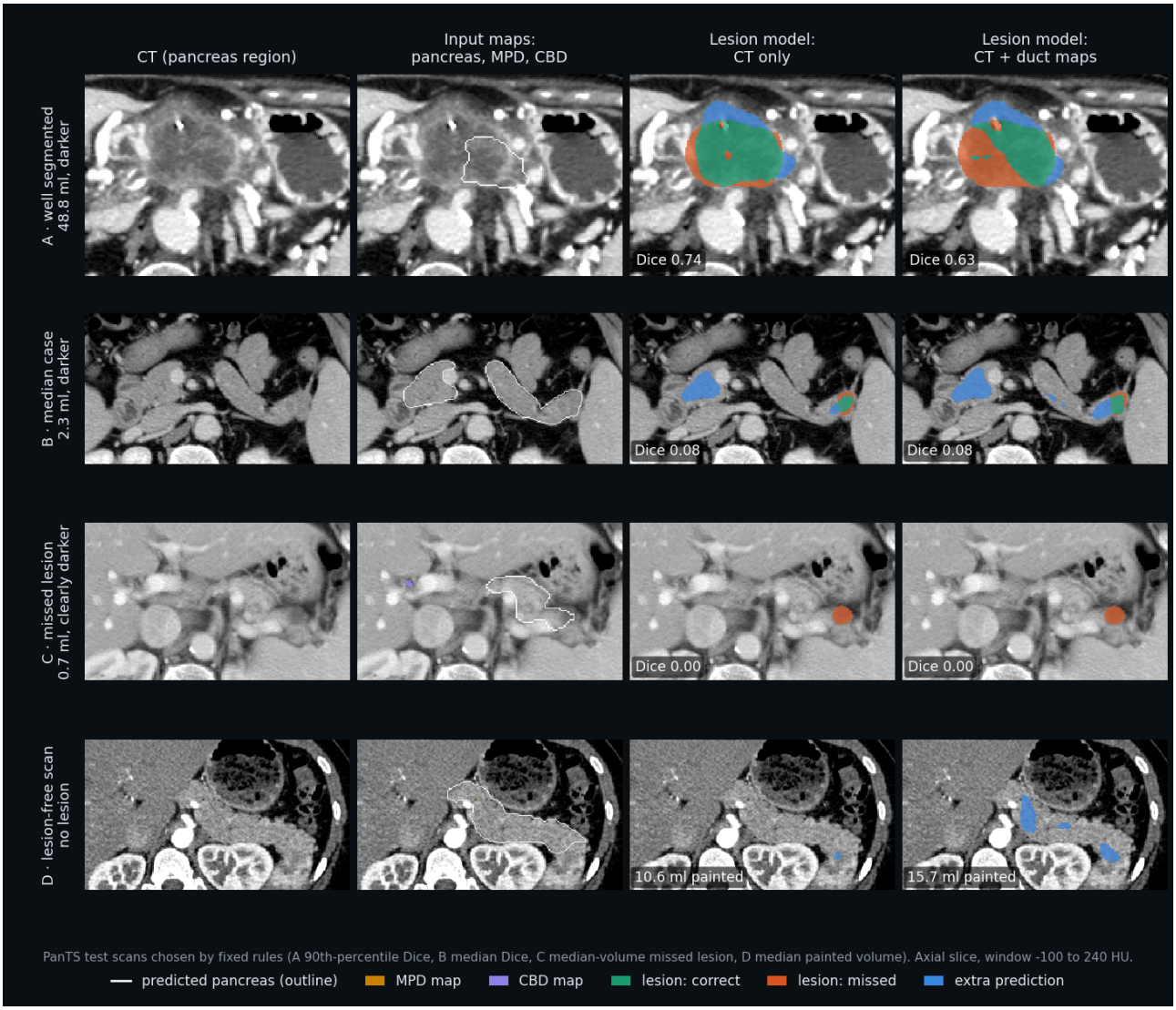


图3。 按固定规则而非凭肉眼选取的四个测试扫描图像：A为最接近采用导管图模型第90百分位Dice系数的扫描，B为最接近其中位Dice系数的扫描，C为漏检病灶中体积处于中位数的病灶，D为预测体积处于中位数的无病灶扫描图像。显示参考病灶（A-C）或预测结果（D）最大的轴位切片，窗宽范围为−100至240亨氏单位。各列依次为：计算机断层扫描；输入图；（预测的胰腺轮廓、主胰管、胆总管）；仅使用计算机断层扫描模型和采用导管图模型的病灶预测（绿色：正确，橙色：漏检，蓝色：额外检出）。每例数值均来自测试评估。

5.2 Dice系数损失发生在哪里（折外）

在折外数据中，病灶平均Dice系数为0.285；168个病灶中有150个被预测结果触及（89.3%），无病灶扫描图像中有预测结果的比例为97.4%。移除所有未触及病灶的连通分量后，Dice系数仅升至0.309。相比之下，如果这些接触病灶的连通分量仅包含病灶，Dice系数将达到0.550。在检出的病例中，接触病灶的连通分量中位体积为12.8 ml，而病灶体积为5.9 ml；其覆盖病灶的中位比例为56.4%，且其体积中有29.0%属于病灶。因此，预测结果既过大又发生了偏移。其非病灶部分位于预测的胰腺内部（中位比例为96.2%），其中46.9%的部分距病灶超过5 mm，且其强度接近正常胰腺（中位差异为−2.2亨氏单位）。提高概率阈值会使预测结果缩小，但会降低Dice系数（阈值为0.5时为0.272，阈值为0.99时为0.236，折0）。

5.3 衰减与大小

表2根据折外病灶相对于正常胰腺的衰减对其进行分层。衰减相近的病灶（168个中有51个）Dice系数最高（0.385）；相近组和较暗组的检出率都很高（分别为94.1%和96.3%）。明显较暗的病灶（44个）中，其体素强度与正常胰腺可区分，这组病灶的Dice系数（0.208）和覆盖率（中位数16.6%）最低。在接触组件内部，在142个可测量病例中，有15个病例的病灶

部分明显比其余部分更暗；其中34个明显较暗的病灶里，有12个在连通成分内部仍明显较暗。小病灶（≤ 2.15 ml）的Dice系数为0.075，中等和大型三分位组分别为0.320和0.460。在测试集上，明显较暗病灶的Dice系数再次最低（0.088（95%置信区间0.047至0.133）），覆盖率也最低（中位数2.1%），且更为常见（127个中有52个）；明显较亮的病灶较少（7个）（表2，右侧）。测试集中明显较暗的病灶体积更小（中位数2.9 ml，折外为5.6 ml），衰减更低（中位数39 HU，折外为50 HU）；28.8%的病灶中位衰减低于30 HU（折外为20.5%）。按体积分组，使用导管图时，测试集在折外小、中、大三分位组对应的Dice系数分别为0.026、0.191和0.428。

表2。 病灶相对于正常胰腺的衰减。折外：胰管图折叠模型。测试：导管图模型。

衰减	折外，n	发现	Dice	覆盖率，中位数	测试，n	测试集 Dice系数（95%置信区间）
明显更亮（曲线下面积 ≤ 0.4）	19	78.9%	0.242	45.4%	7	0.352（95%置信区间 0.167 至 0.547）
相似（0.4–0.6）	51	94.1%	0.385	68.3%	25	0.333（95%置信区间 0.242 至 0.430）
较暗（0.6–0.8）	54	96.3%	0.269	46.7%	43	0.271（95%置信区间 0.193 至 0.358）
明显更暗（> 0.8）	44	79.5%	0.208	16.6%	52	0.088（95%置信区间 0.047 至 0.133）

图4展示了两种病灶模型在测试集上的 Dice 系数，并按衰减和病灶大小进行分层。

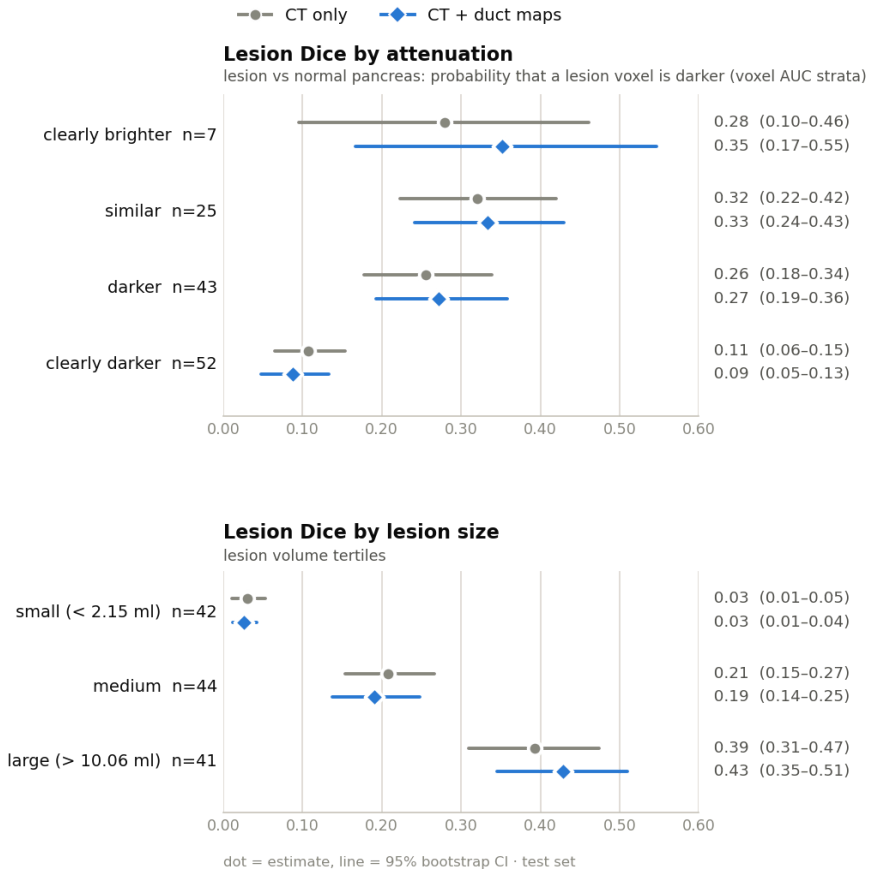


图4。按衰减分层（病灶与正常胰腺的体素 AUC）和病灶体积分位组划分的测试集病灶 Dice系数，适用于两种病灶模型（点：估计值；线：95%病例自助法置信区间）。

5.4 候选筛选

在使召回80%病灶扫描的工作点上，仍有候选项的无病灶扫描占比分别为：裁剪卷积神经网络（crop CNN）87.4%、连通分量体积79.5%，以及梯度提升74.5%（表 3）。最大候选项得分的扫描级 AUC分别为0.631、0.674和0.716。在连通分量级上，三种筛选条件在召回率为80%时的精确率均较低（表 3）。裁剪卷积神经网络并未优于单一体积特征。

表 3。候选筛选，折外（168 个病灶扫描，1433 个无病灶扫描）。

筛选条件	含候选项的无病灶扫描在 80% 找到	扫描级 AUC	召回率为 80% 时的连通成分精确率（第 1 折 / 第 0 折候选项）
连通分量体积	79.5%	0.674	0.078 / 0.108
梯度提升，30 测量值	74.5%	0.716	0.070 / 0.074
CT裁剪区域上的 3D ResNet-10	87.4%	0.631	0.067 / 0.065

5.5 导管和胆总管直径作为单独征象

表4和图2最下方的面板显示，仅凭预测图区分病灶扫描与无病灶扫描的效果。主胰管直径的测试集AUC达到0.794（95%置信区间0.747至0.839）（折外AUC为0.665（95%置信区间0.613至0.715））。仅使用胆总管直径的效果较弱（0.603（95%置信区间0.549至0.658））；将其加入主胰管后，测试集AUC没有变化（差异为-0.000（95%置信区间-0.009至+0.010））；胆总管的拟合权重较小（标准化系数：主胰管为0.759，胆总管为0.148）。主胰管信号有一部分来自病灶本身的图像体素：在24.4%的测试病灶扫描中，至少10%的预测主胰管位于病灶内部；移除这些体素后，测试集AUC降至0.758（95%置信区间0.708至0.806）。病灶扫描中的主胰管直径中位数为4.8 mm，而测试无病灶扫描中为3.2 mm。按病灶大小分层，主胰管直径的测试集AUC在小、中、大病灶中分别为0.730、0.878和0.767（折外AUC分别为0.601、0.659和0.735）。折外预测图来自在这些扫描图像上训练的导管模型；折外AUC与测试集AUC存在差异，但我们未探究其原因。

表4。仅根据预测图判断扫描级病灶存在情况（最大直径）。MPD + CBD：在折外扫描图像上拟合的逻辑回归；其折外AUC为样本内AUC。

Sign	折外AUC（95% CI)	测试集AUC（95%置信区间）	测试集AUC，病灶体素已移除	按病灶大小划分的测试 AUC（小 / 中型/大型）
仅主胰管	0.665（95%置信区间 0.613 至 0.715）	0.794（95%置信区间 0.747 至 0.839）	0.758	0.730 / 0.878 / 0.767
仅胆总管	0.580（95%置信区间 0.530 至 0.630）	0.603（95%置信区间 0.549 至 0.658）	0.598	0.624 / 0.578 / 0.610
主胰管 + CBD	0.673（95%置信区间 0.620 至 0.721）	0.793（95%置信区间 0.746 至 0.839）	0.758	0.730 / 0.876 / 0.770

表5给出了固定截断值。主胰管直径至少为5 mm时具有特异性（检验特异度为95.9%），但仅见于少数病灶扫描图像中（检验敏感度为44.1%）。许多无病灶扫描图像中预测的胆总管较宽（中位数为6.9 mm），因此不存在兼具敏感度和特异度的胆总管截断值。

表5。固定直径截断值的敏感度和特异度（折外包括168例病灶扫描图像和1422例无病灶扫描图像；测试集包括127例和608例）。

截断值	折外敏感度	折外特异度	测试敏感度（95%置信区间）	检验特异度（95%置信区间）
主胰管 ≥ 3 毫米	82.7%	23.4%	96.1%（95%置信区间 92.1%–99.2%）	25.8%（95%置信区间 22.5%–29.4%）
主胰管 ≥ 4 毫米	57.1%	68.1%	66.1%（95%置信区间 58.3%–74.0%）	73.4%（95%置信区间 70.1%–77.0%）
主胰管 ≥ 5 毫米	33.3%	95.2%	44.1%（95%置信区间 35.4%–52.8%）	95.9%（95%置信区间 94.2%–97.4%）
主胰管 ≥ 6 mm	22.6%	98.0%	33.9%（95%置信区间 26.0%–41.7%）	97.4%（95%置信区间 96.1%–98.5%）
胆总管 ≥ 6 mm	81.5%	21.4%	86.6%（95%置信区间 80.3%–92.1%）	24.0%（95%置信区间 20.7%–27.6%）
胆总管 ≥ 8 mm	44.6%	67.2%	41.7%（95%置信区间 33.1%–50.4%）	72.9%（95%置信区间 69.2%–76.5%）
胆总管 ≥ 10 mm	23.2%	92.3%	16.5%（95%置信区间 10.2%–23.6%）	93.6%（95%置信区间 91.4%–95.6%）
胆总管 ≥ 12 毫米	10.1%	97.7%	6.3%（95%置信区间 2.4%–11.0%）	98.2%（95%置信区间 97.0%–99.2%）

6. 讨论

这一分解结果表明，病灶扫描图像中的假阳性连通区域并非Dice系数偏低的主要原因：即使连通域滤波器完美无缺，提升也微乎其微；而定位正确的连通成分范围过大，造成了大部分性能损失。多出的部分是外观正常的胰腺组织，且往往距离病灶数毫米，这解释了为何提高阈值或使用连通域级分类器都无济于事。

衰减分析揭示了两个问题。约三分之一的病灶与正常胰腺具有相同的衰减；对于这些病灶，计算机断层扫描中几乎没有边界信息，因此任何基于强度的方法都难以获得较高的重叠度，这在预料之中。相比之下，明显较暗的病灶在强度上清晰可见，但无论是在折外数据还是测试集上，它们的覆盖率都最低。因此，图像中存在模型未能利用的边界信息。我们无法判断这是学习问题还是病灶类型所致：PanTS 未区分囊肿与实性病灶，而其中部分病灶的衰减接近水。这是未来研究中一个明确的目标。

像 [3]，中那样将预测的胰管图作为额外输入，并未改变测试集上的病灶Dice系数。它们使病灶模型总体上预测得更多：更多病灶被覆盖，无病灶扫描也获得了更大的预测区域，因此更高的检出率并未改善扫描级区分。

几乎所有无病灶扫描中都出现预测，这是将此流程用于扫描级检测的主要障碍。我们尚未确定这是由训练设置（例如，无病灶扫描对批次Dice损失的贡献有多小）还是由数据本身造成的；这仍是一个有待解决的问题。

局限性。PanTS中病灶类型未知，因此这些结果混合了胰腺导管腺癌、其他病灶和囊肿。数据集按扫描图像划分，未使用患者标识符。折外分析使用了五折中的两折，而这些扫描图像在第1和第2阶段属于样本内数据。该测试集此前曾用于比较胰管模型。导管和胆总管直径分析是在主要测试运行后确定的；其中唯一拟合的部分使用折外扫描图像进行训练。测试集Dice系数低于验证集Dice系数（使用导管图时为0.276，涉及16个病灶），也低于折外Dice系数；测试扫描图像来自PanTS ImageTe分区，训练扫描图像来自ImageTr。两个分区的病例构成有所不同：静脉期扫描图像占测试集的77.3%，而训练集为50.0%（病灶扫描图像中的比例分别为79.5%和56.6%），且明显较暗的病灶更常见；两者是否也存在来源机构差异尚未进行考察。本研究没有外部测试集，也没有阅片者研究。胰管图的益处仅针对一种架构和一种训练时长进行了评估。

7. 结论

在一个大型公开CT数据集上，端到端胰腺 → 导管 → 病灶流程的测试集病灶Dice系数为0.213；将预测的胰管图作为额外输入并未改变这一结果。损失主要来自定位正确的预测范围有限，以及无病灶扫描图像上的预测。等衰减病灶限制了基于强度的边界所能达到的效果，而明显的低衰减病灶在折外和测试集中覆盖率都最低，却具有模型未能利用的边界，是最明确的改进目标。

代码和数据可用性

预处理、训练、推理和评估的代码，以及数据划分，均可在 https://github.com/Kang812/pants-lesion-failure-analysis 获取。在八个测试扫描图像上重新运行已发布的评估脚本，得到的每个病例数值与保存的结果完全一致。PanTS 遵循其自身许可协议（CC BY-NC-SA 4.0）。

参考文献

1. Li W 等。基于计算机断层扫描（ePAI）早期及诊断前的胰腺癌检测。arXiv:2601.22134，2026。 2. Alves N 等。用于计算机断层扫描胰腺导管腺癌检测的全自动深度学习框架。Cancers 2022。 doi:10.3390/cancers14020376。 3. Viviers CGA、Ramaekers M 等。利用临床相关次要特征改进胰腺肿瘤检测。arXiv:2208.03581，2022。 4. Yamaguchi T、Sofue K 等。深度学习检测增强CT和平扫CT上与胰腺癌相关的直接和间接影像学征象。Radiology 2026。 doi:10.1148/radiol.253122。（仅阅读摘要。） 5. Abi Nader C、Vetil R 等。利用深度学习自动检测门静脉期CT扫描上的胰腺病灶和主胰管扩张。Invest Radiol 2023。 doi:10.1097/rli.0000000000000992。（仅阅读摘要。） 6. Li W、Zhou X、Chen Q 等。PanTS：胰腺肿瘤分割数据集。arXiv:2507.01291，2025。 7. Myronenko A。利用自编码器正则化进行三维磁共振成像脑肿瘤分割。arXiv:1810.11654，2018。 8. Cardoso MJ、Li W、Brown R 等。MONAI：用于医疗保健深度学习的开源框架。arXiv:2211.02701，2022。