类别不平衡校正以改善 CT 中的通用病变检测和标记

Peter D. Erickson, Tejas Sudharshan Mathai*, Ronald M. Summers

Imaging Biomarkers and Computer-Aided Diagnosis Laboratory, Radiology and Imaging Sciences, Clinical Center, National Institutes of Health, Bethesda, MD, USA
* Corresponding author, email: mathaits [at] nih dot gov

摘要 放射科医生通常会在 CT 扫描中检测和测量病灶，以评估癌症的分期和肿瘤负担。为了可能帮助他们的工作，开发了多个病灶检测算法，并使用了一个名为 DeepLesion（包含 32,735 个病灶、32,120 张 CT 切片、10,594 项研究、4,427 名患者、8 种身体部位标签）的大型公共数据集。然而，该数据集中存在缺失的测量值和病灶标签，并且在每个标签类别中的病灶数量不平衡问题非常严重。在这项工作中，我们利用 DeepLesion 的一个有限子集（6%，1331 个病灶、1309 张切片）进行训练 VFNet 模型以检测病灶并对其进行标记。该子集中包含病灶注释和身体部位标签。为了解决类别不平衡问题，我们进行了三项实验：1) 通过身体部位标签平衡数据；2) 通过每位患者中的病灶数量平衡数据；3) 通过病灶大小平衡数据。与随机采样的（不平衡的）子集相比，我们的结果显示，通过平衡身体部位标签总能增加低数据量类别的病灶敏感性（小于 ≥1cm：骨骼:80%对 46%，肾脏:77%对 61%，软组织:70%对 60%，骨盆:83%对 76%）。其他三种测试模型（FasterRCNN、RetinaNet、FoveaBox）也显示了类似的趋势。通过病灶大小平衡数据也有助于 VFNet 模型在对比不平衡的数据集时改善所有类别的召回率。我们还提供了一种结构化的报告指南，用于将“病灶”子部分输入放射科报告的“发现”部分中。据我们所知，我们是第一个报告 DeepLesion 中的类别不平衡问题，并采取了数据驱动的步骤来解决在联合病灶检测和标记上下文中该问题。

Keywords: , 全局病变检测, 类别不平衡, 深度学习, DeepLesion

1 介绍

肿瘤负荷评估和癌症分期对于患者治疗 [1, 2] 至关重要。实现这一目标的第一步是病灶定位，这使得能够测量病灶大小并评估恶性风险。通常，在临床实践中，计算机断层扫描（CT）和正电子发射断层扫描（PET）被优选用于病变分析 [1]。放射科医生滚动浏览整个体积以找到尺寸为 ≥10 毫米

的病灶，并将其视为可疑转移病灶 [1, 2]。他们还跨多个就诊识别病灶并追踪其进展（增长、缩小或不变），基于治疗反应。CT 中的病变可能具有异质形状、大小和外观，这进一步使评估复杂化，因为存在多种成像扫描仪和不同机构使用的不一致检查协议。此外，在繁忙的临床日中对病变进行测量对于放射科医生来说是困难的，由于观察者测量变异性，特别是当检查转移的治疗指南演变时，某些潜在的转移性病灶可能会被遗漏。

最近，许多自动化方法被提出用于在 DeepLesion 数据集上进行通用病变检测 [3–9]，并取得了最先进的结果。DeepLesion 数据集包含了放射科医生标注的 32,735 个病灶，这些病灶来自 4,427 名患者 10,594 项研究中的 32,120 个轴向 CT 切片。该数据集被划分为 70% 训练集、15% 验证集和 15% 测试集。八个（8）病灶级标签（骨骼、腹部、纵隔、肝脏、肺部、肾脏、软组织和骨盆）仅在验证集和测试集中可用。先前的研究利用整个数据集进行开发和测试，而只有少数研究超越了病变检测并解决了临床问题 [8–11]。然而，如图 1（a）所示，DeepLesion 数据集中存在严重的类别不平衡现象，某些类别的病灶（肺部、腹部、纵隔和肝脏）过度表示，而其他较少见的类别（骨盆、软组织、肾脏、骨骼）则代表性不足。这种不平衡在先前的工作中尚未得到解决；例如，在 [3] 中，为了提高检测性能，增加了用于肺结节（LUNA 数据集 [12]）、肝肿瘤（LITS 数据集 [13]）和淋巴结（NIH 淋巴结数据集 [14]）的公共数据集。然而，这仅仅增加了代表性过多类别中的数据量（及检测性能），而没有影响代表性不足的类别。解决类不平衡具有潜在的临床意义，例如提高时间间隔变化检测（随时间病变跟踪）[8, 9]。

在本文中，我们通过仅使用标注子集（30%）来训练最先进的 VFNet 模型 [15] 以解决 DeepLesion 数据集中类别不平衡的问题，并进行病变检测和分类。在有限的数据驱动方式下，我们根据以下内容平衡了训练数据：1) 病变被识别的部位，2) 患者中观察到的病变数量，3) 病变的大小。通过按身体部位标签来平衡数据，我们显示了对代表性不足（UR）类别的检测灵敏度的一致性增加以及对过度代表（OR）类别灵敏度的最小下降。这种趋势在其他检测器中也可以看到，例如 Faster RCNN[16]、RetinaNet[17] 和 FoveaBox[18]。此外，我们还通过根据病变大小平衡实验中的病变，观察到了 VFNet 模型对所有类别的召回率提升。另外，我们通过创建一个专门的“病变”子部分来提供结构化的报告指南，该子部分用于放射学报告中“发现”部分的输入。“病变”子部分包含检测到的病变的结构化列表以及它们的身体部位标签、检测置信度和系列及切片号。据我们所知，我们是第一个

展示 DeepLesion 数据集中的类别不平衡问题，并且在病变检测和分类背景下采取了基于数据驱动的方法来解决这一问题的研究。

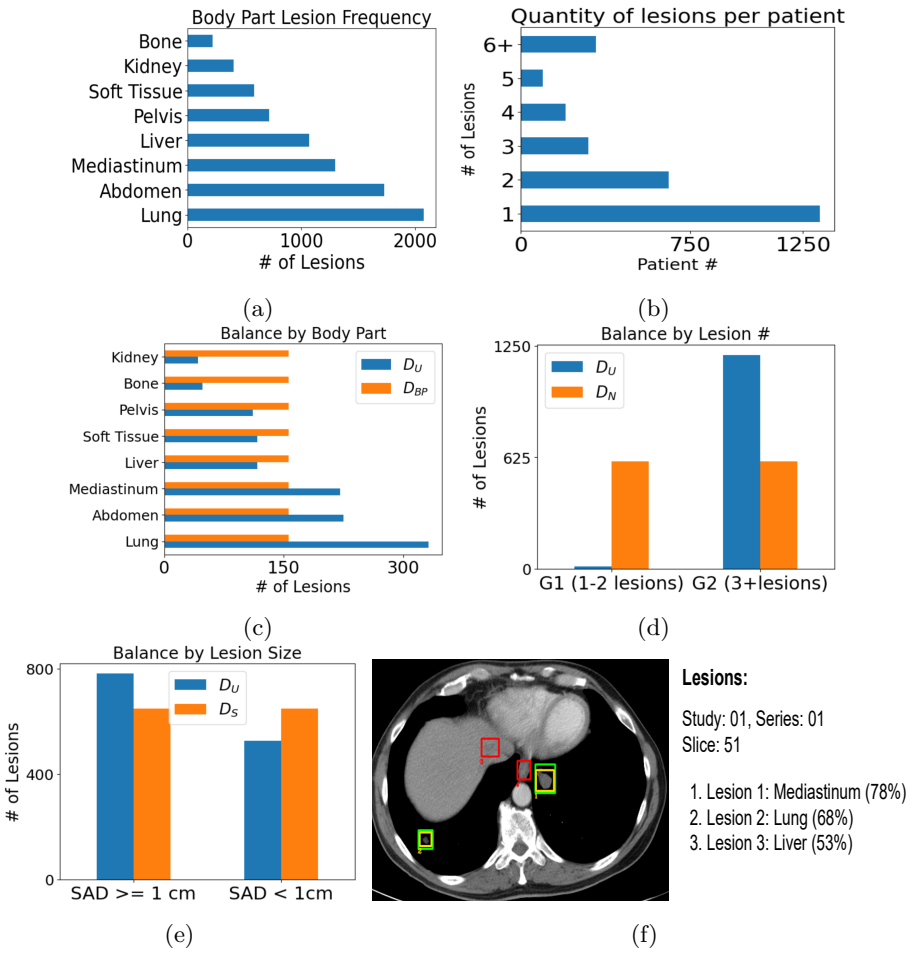


图 1: (a) 显示了 DeepLesion 数据集 [22] 中每个身体部位标签的病灶分布，某些类别过度或不足。(b) 显示了具有特定数量标注病灶的患者数量。(c) 与不平衡的数据集 D_U 相比，我们的数据集 D_{BP} 在不同身体部位类别的病灶数量上实现了平衡（橙色）。(d) 展示了将患者分为两组后的病灶分布：G1 组的患者有 1-2 个病灶，而 G2 组的患者有 3 个或更多病灶。与 D_U 相比，数据集 D_N（橙色）在 G1 和 G2 中具有相等数量的病灶。每个组中的患者数量并未平衡。(e) 显示了按短轴直径 (SAD) 长度分类的病灶分布。与 D_U 相比，在数据集 D_S 中，SAD≥1cm 和 SAD<1cm 的病灶数量是平衡的（橙色）。(f) 在胸部区域检测到四个病灶。绿色框：GT，黄色框：TP，红色框：FP。前三个预测结果及其标签和置信分数被编译成一个结构化的 “病灶” 子部分，用于录入放射学报告的 “发现” 部分。仅显示预测置信度 ≥50%的病灶。最佳电子彩色查看图形。

2 方法

在本节中，我们简要描述了用于病灶检测和标记的神经网络。我们的目标是通过数据驱动的方法提高现有模型对类别不平衡的鲁棒性。

最先进的检测器 各种最先进的检测器被用于 CT 切片中的病灶检测和标记。值得注意的是，我们使用了：1) VFNet[15]，2) Faster RCNN[16]，3) RetinaNet[17] 和 4) FoveaBox[18]。Faster RCNN 是一种基于锚点的两阶段检测器，其中第一阶段生成感兴趣区域（ROI）的区域建议，随后第二阶段对这些建议进行分类并回归边界框坐标。RetinaNet[17] 是一种无锚检测器，利用焦点损失来解决检测中常见的类别不平衡问题，在该问题中，建议框是在图像的非信息区域而不是显著物体位置进行采样的。FoveaBox 使用 ResNet-50 主干网从输入生成特征图，并使用了一个估计图像中可能被对象 ROI 覆盖坐标的黄斑头部网络。最后，VFNet 结合了全卷积单阶段目标检测器 FCOS [19]（不包括中心度分支）、自适应训练样本选择机制 ATSS [20]，该机制在训练过程中选择了高质量的 ROI 候选区域，以及一种新的感知 IoU 的变焦点损失 [15] 来检测 ROI。模型训练完成后，加权框融合 WBF [21] 被用来结合众多预测并提高精度/召回率指标。补充材料包含模型的实现细节。

3 实验

数据集。NIH DeepLesion 数据集 [22] 包含了带有每片 1-3 处病灶注释的关键切片，并提供了关键切片上下各 30mm 的背景。注释是使用 RECIST 测量方法 [2] 进行的，从中为每个病灶提取了二维边界框。八个（8）个病灶级别的标签（骨骼 - 1，腹部 - 2，纵隔 - 3，肝脏 - 4，肺部 - 5，肾脏 - 6，软组织 - 7，骨盆 - 8）仅可用于验证和测试数据集的分割部分。病灶标签是通过一个身体部位回归器 [23] 获得的，该回归器提供了一个连续得分，表示了切片在 CT 图像体积中的标准化位置（例如肝脏、肺部、肾脏等）。任何标注有病变的切片被分配给相应切片的身体部分标签 [22, 23]。由于 DeepLesion 数据集包含了同一患者的多次就诊记录，仅保留了第一次就诊时的病灶 [8]，以保持训练过程中的唯一性。这一过程使数据集中剩下来自 25,568 张切片的 26,034 处病变。接下来，我们移除了没有包含身体部位标签的病变（训练分割），使我们剩下了一个有限的数据集 D_L，其中包含了来自 7,953 个切片的 8,104 个病变。D_L 包含了原始 DeepLesion 数据集 ∼24.75%，并被分割为 60%训练、

20%验证和 20%测试，每个分割中包含独特的患者。这个 60%的训练分割仍然不平衡，在我们的实验（见下文）中，我们只使用了 DeepLesion~6%的数据（1,331 个病变，1,309 个切片）。

实验设计。图 1(a) 中每个身体部位标签的不均衡病灶分布以及图 1(b) 中每位患者的病灶数量分布促使我们设计了四个实验，使用了一个有限标注数据集 D_L。在第一次实验 E_{BP} 中，我们生成了一个通过身体部位标签平衡的数据集 D_{BP}；如图 1(a) 和 1(c) 所示，数据量最少的身体部位标签（“骨骼”）被识别出来，并且其他标签中的数据量与最低数量相匹配。目的是通过样本选择强调在训练过程中所有病灶类别都是等可能的。在第二次实验 E_N 中，我们创建了一个根据患者拥有的病灶数量平衡的数据集 D_N；从图 1(b) 可以看出，有大量患者拥有 1-2 个病灶，而拥有 3 个或更多病灶的患者较少。对于 E_N，我们首先创建了两个组（G1 和 G2），并为每个组采样患者，使得每组包含的病灶数量如图 1(d) 所示相同。目的是提供每位患者的病灶数量平衡，以便模型在测试时以相同的概率见证具有不同数量病灶的患者。我们的第三次实验 E_S 具有临床导向，并生成了一个按病灶大小 D_S 平衡的数据集。SAD≥10mm 的病灶被收集在一个组中，而那些 SAD<10mm 的则存在于第二个组中。目的是创建一个根据其尺寸划分的病灶数量相等的数据集，因为在测试时小病灶和大病灶出现的概率是相同的。在我们的第四次也是最后一次实验 E_U 中，我们生成了一个不平衡的数据集 D_U，它采用了训练分割 D_L 的随机样本，使得标签的分布（随机）类似于图 1(c)-(e) 所示。

4 结果与讨论

结果。表 1和图 2显示了我们对 SAD≥ 大于 1 厘米的病灶在 4 FP 和 30% IoU 重叠 [24] 的实验结果，这些病灶通常更具临床意义。补充材料中的表 1 反映了我们对 SAD 小于 1 厘米的病灶的实验结果。与不平衡数据的实验 D_U 相比，我们的平衡身体部位标签的实验 E_{BP} 对于所有测试模型中的 4/4 个代表性不足（UR）类别（骨：80%对 46%，肾：77%对 61%，软组织：70%对 60%，骨盆：83%对 76%）提高了召回率。这些结果对于 SAD≥ 和 $<$ 1cm 都很明显。在代表性过高的（OR）类别中，我们看到所有模型的“肺”类别表现一致改善，并且除了 FoveaBox 所有模型中的“纵隔”标签也有所改进。然而，“腹部”和“肝脏”类别的敏感性在所有模型中都有所下降。这种敏感性的降低是预期之内的，因为用于 OR 类别的训练样本数量已经减少如图 1(c) 所示。尽管如此，为了更好地理解这一现象，我们使用 VFNet 模型

计算了每个实验的混淆矩阵（见补充材料）。从 DeepLesion 数据集描述 [22] 可知，“软组织” 类别涵盖了在肌肉、皮肤或脂肪中发现的所有病变，而 “腹部” 类别是一个涵盖所有非 “肾脏” 或 “肝脏” 的腹部病变的通用术语。然而，解剖学上，“肾脏” 和 “肝脏” 是彼此相邻的器官，轴向切片经常在同一张图片中显示这两个器官的横截面。这一点在混淆矩阵中得到了体现，即 “腹部”、“肾” 和 “肝” 标签之间最常被混淆。比较 E_U 和 E_N（按病变数量平衡），对于 Faster RCNN 和 FoveaBox 只有 1/4 UR 类别（肾脏）的召回率有所改善，而 RetinaNet 分别为 2/4 UR 类别（肾脏和软组织）以及 VFNet（肾脏、软组织）。对于 OR 类别，只有 VFNet 的 “肝脏” 得到改善，Faster RCNN 和 FoveaBox 分别有 2/4 类别有所改进（腹部，肺），而 RetinaNet 则有 3/4 类别得到提升（纵隔，肺，肝）。与 E_{BP} 相比，所有 UR 类别的召回率都较低，除了 VFNet，在 2/4 类别（软组织和骨盆）上表现良好。对于 OR 类别，Faster RCNN 在 2/4 类别中有所改善（腹部、肝脏），VFNet 和 RetinaNet 在 3/4 类别中有所改善（腹部、纵隔和肝脏），而 FoveaBox 在所有 4 个类别中都有所改善。

与 E_U 实验相比，在 E_S 实验（通过病灶大小平衡）中，VFNet 的召回率始终在所有类别中都更好，无论是 SAD≥1cm 还是 $<$ 1cm。只有一个 UR 类别的召回率有所提高，分别是 RetinaNet 和 FoveaBox (Bone)，而 3/4 UR 类别对 Faster RCNN 的表现较好（Bone, Kidney, Soft Tissue）。在 OR 类别中，RetinaNet 在所有类别上的表现都较差，并且有 2/4 OR 类别的表现有所提高，分别是 Faster RCNN 和 FoveaBox (Abdomen, Lung)。与 E_{BP} 实验相比，除 “Pelvis” 类之外，VFNet 对所有 UR 类别都具有较低的灵敏度。对于 OR 类别，RetinaNet 没有任何类别的表现得到改善，2/4 类别对 Faster RCNN 的表现有所提高（Abdomen, Liver），3/4 类别对 FoveaBox 的表现有所提高（Abdomen, Mediastinum, Liver）。与 E_N 实验相比，RetinaNet 对所有 UR 类别的灵敏度都较差。FoveaBox 在 2/4 UR 类别中表现更好（Bone 和 Pelvis），而 Faster RCNN 在 3/4 UR 类别中表现更好（Bone, Kidney, Soft Tissue）。在 OR 类别上，RetinaNet 的召回率对所有类别都较差，1/4 OR 类别的表现有所提高，Faster RCNN (Lung)，VFNet 对 3/4 类别的表现有所提高（Abdomen, Mediastinum, Lung），而 FoveaBox 对所有类别的表现都有所提高。

讨论。与之前的工作相比，我们已经证明通过有效探索 DeepLesion 数据集，所有模型在所有代表性不足类别上的召回率都有所提高。具体来说，我们的

表 1: 不同探测器基于不同实验的检测灵敏度在病变 SAD 为 ≥1 厘米时，@4FP 和 IOU 为 0.3 的情况下显示。

Experiment	Bone	Kidney	Soft Tissue	Pelvis	Abdomen	Mediastinum	Lung	Liv
E_U - Faster R-CNN [16]	23.3	40.5	50.4	67.5	57.7	79.6	66.8	**77**
E_{BP} - Faster R-CNN [16]	**63.3**	**75.1**	**58.1**	**68.5**	55.6	**83.3**	**74.9**	69
E_N - Faster R-CNN [16]	16.6	49.3	44.1	63.2	**65.5**	78.7	72.6	76
E_S - Faster R-CNN [16]	30.0	49.7	51.7	56.4	61.1	74.8	74.5	73
E_U - RetinaNet [17]	21.7	38.4	48.2	55.8	**70.5**	82.8	76.1	75
E_{BP} - RetinaNet [17]	**66.7**	**66.7**	**60.3**	**59.8**	62.3	85.3	**79.7**	71
E_N - RetinaNet [17]	27.5	53.1	26.1	49.6	68.7	**86.2**	77.4	**76**
E_S - RetinaNet [17]	26.2	22.4	25.0	21.1	51.9	61.7	58.8	58
E_U - FoveaBox [18]	28.3	46.4	54.2	59.2	64.8	**88.3**	69.2	**76**
E_{BP} - FoveaBox [18]	**65.0**	**67.9**	**66.7**	**63.4**	56.2	84.5	**76.9**	70
E_N - FoveaBox [18]	18.3	56.5	46.9	34.1	70.1	85.1	74.7	71
E_S - FoveaBox [18]	41.66	40.9	46.3	47.6	**71.1**	86.8	75.1	74
E_U - VFNet [15]	46.7	61.6	60.0	76.0	76.8	85.6	70.8	77
E_{BP} - VFNet [15]	**80.0**	**77.6**	**70.7**	83.4	69.5	87.7	78.9	76
E_N - VFNet [15]	28.8	63.3	63.6	73.5	69.6	78.8	68.2	**91**
E_S - VFNet [15]	51.6	67.0	67.3	**87.2**	**82.1**	**89.8**	**82.7**	82

E_{BP} 实验（通过身体部位标签平衡数据）显示了这一明显的改进。我们也看到，在 “Lung” 和 “Mediastinum” 类中，Faster RCNN、RetinaNet 和 VFNet 的敏感性有所增加。“Abdomen” 和 “Liver” 类别经常被混淆。我们主张，“Abdomen” 和 “Soft Tissue” 标签是通过身体部位回归器生成的，这些标签具有模糊性和非特异性，并广泛涵盖了腹部多个区域。事实上，在请放射科医生对 100 个随机样本病变使用原始术语 “Abdomen” 重新分类后，我们发现了多个应被分配新标签如 “Liver”、“Pancreas”、“Spleen”、“Muscle”、“Stomach” 等的病变。在 DeepLesion 数据集中有许多其他标注的病变，在手动检查时其分配的标签可能会发生变化。在我们的 E_N 实验（按病灶数量平衡）中，我们没有看到一致的改进趋势，并假设这是因为未同时通过身体部位标签来平衡病灶。同时按照身体部位标签和病灶数量来平衡数据证明很困难，因为当患者有多个带有不同标签的病变时很难进行分类。在我们的

E_S 实验（平衡病灶大小）中，所有类别的召回率都随着 VFNet 模型的使用而提高。

我们无法与先前的工作进行比较，因为联合检测和标记病变 [10, 11] 的方法有限。一种方法 [11] 使用了需要分割标签的 Mask-RCNN 模型，在这项工作中我们并未创建这些标签。此外，这种方法还提供了更详细的描述性标签，这需要从放射学报告中得出一套复杂的本体论来将它们映射到本工作中使用的部位标签（在 DeepLesion 数据集中不可用）。为解决此问题，我们实现了其他检测模型以证明我们的结果具有一致性。此外，我们通过创建一个专门的“病变”子部分进入放射学报告中的“发现”部分，呈现了一个有用的临床结构化报告指南。“病变”子部分包含一个带有其部位标签、检测置信度以及序列和切片号的结构化病变列表。另外，DeepLesion 数据集同时包含了对比增强 CT 体层和非对比增强 CT 体层，但具体相位信息在该数据集中不可用。因此，我们无法根据 CT 体层的相位来平衡数据，这是我们工作的局限性。未来的工作计划进行一个实验，对低样本点类别的上采样，并通过部位标签和病变大小来平衡数据。

5 结论

在本文中，我们展示了 DeepLesion 数据集在每个身体部位标签的病灶数量上存在严重的不平衡。它还包含缺失的注释和标签标记。我们利用了有限的数据子集（6%，1331 个病灶，1309 个切片）来训练 VFNet 模型以检测病灶并对其进行标记。我们进行了三项实验来解决类别不平衡问题，并通过我们的实验 E_{BP} 展示了 UR 标签召回率的一致增加（骨：80%对 46%，肾：77%对 61%，软组织：70%对 60%，骨盆：83%对 76%），与 E_U 相比。我们还表明，FasterRCNN、RetinaNet 和 FoveaBox 表现相似。此外，我们展示了通过病灶大小平衡数据有助于 VFNet 模型提高所有类别的召回率。据我们所知，我们是第一个展示 DeepLesion 数据集中类别不平衡问题的，并在病灶检测和分类背景下采取了基于数据驱动的方法来解决这一问题。

致谢。本工作得到了美国国家卫生研究院（NIH）临床中心内部研究计划的支持。

参考文献

1. E. Eisenhauer et al., “New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1)”, Eur J Cancer, 45(2), pp. 228–247 (2009).

2. L. Schwartz et al., “Recist 1.1—update and clarification: From the recist committee”, European J Cancer, 62, pp. 132–137 (2016).

3. K. Yan et al., “Learning from Multiple Datasets with Heterogeneous and Partial Labels for Universal Lesion Detection in CT”, In: IEEE TMI, 40(10), pp. 2759–2770 (2021).

4. J. Cai et al., “Lesion Harvester: Iteratively Mining Unlabeled Lesions and Hard-Negative Examples at Scale”, In: IEEE TMI, 40(1), pp. 59–70 (2021).

5. J. Yang et al., “AlignShift: Bridging the Gap of Imaging Thickness in 3D Anisotropic Volumes”, In: MICCAI, pp. 562–572 (2020).

6. J. Yang et al., “Asymmetric 3D Context Fusion for Universal Lesion Detection”, In: MICCAI, (2021).

7. L. Han et al., “SATr: Slice Attention with Transformer for Universal Lesion Detection”, In: arXiv, (2022).

8. J. Cai et al., “Deep Lesion Tracker: Monitoring Lesions in 4D Longitudinal Imaging Studies”, In: IEEE CVPR, (2020).

9. W. Tang et al., “Transformer Lesion Tracker”, In: arXiv, (2022).

10. K. Yan et al., “Holistic and Comprehensive Annotation of Clinically Significant Findings on Diverse CT Images: Learning from Radiology Reports and Label Ontology”, In: IEEE CVPR, (2019).

11. K. Yan et al., “MULAN: Multitask Universal Lesion Analysis Network for Joint Lesion Detection, Tagging, and Segmentation”, In: MICCAI, (2019).

12. A. A. A. Setio et al., “Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: The LUNA16 challenge”, Med. Image Anal., 42, pp. 1- – 13 (2017).

13. P. Bilic et al., "The Liver Tumor Segmentation Benchmark (LiTS)”, In: CoRR, (2019).

14. H. Roth et al., "A New 2.5D Representation for Lymph Node Detection Using Random Sets of Deep Convolutional Neural Network Observations”, In: MICCAI, (2014).

15. H. Zhang et al., “VarifocalNet: An IoU-Aware Dense Object Detector”, In: IEEE CVPR, pp. 8514–8523 (2021).

16. S. Ren et al., “Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks”, In: IEEE PAMI, 39(6), pp. 1137–1149 (2017).

17. T.Y. Lin et al., “Focal Loss for Dense Object Detection”, In: IEEE ICCV, pp. 2999–3007 (2017).

18. T. Kong et al., “FoveaBox: Beyond Anchor-based Object Detector”, In: arXiv, (2019).

19. Z. Tian et al., “FCOS: Fully Convolutional One-Stage Object Detection”, In: IEEE ICCV, pp. 9627–9636 (2019).

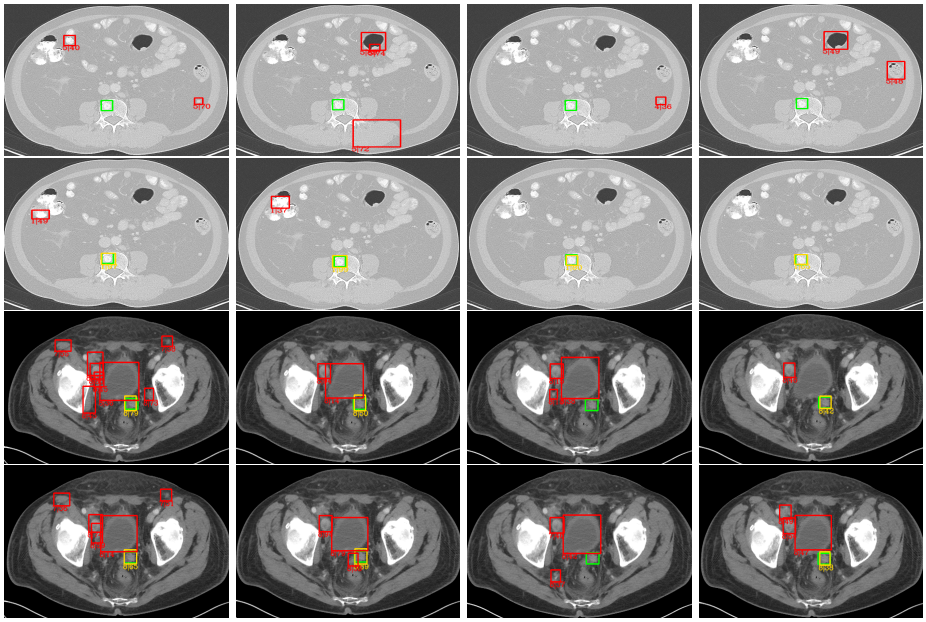
20. S. Zhang et al., “Bridging the Gap Between Anchor-based and Anchor-free Detection via Adaptive Training Sample Selection”, In: CoRR, (2019).

21. R. Solovyev et al., “Weighted Boxes Fusion: Ensembling Boxes from Different Object Detection Models”, Img. Vis. Comp., 107, (2021).

22. K. Yan et al., “DeepLesion: automated mining of large-scale lesion annotations and universal lesion detection with deep learning”, J Medical Imaging, 5(3), pp. 1–11 (2018).

23. K. Yan et al., “Unsupervised body part regression via spatially self-ordering convolutional neural networks”, In: IEEE ISBI, pp. 1022–1025, (2018).

24. T. Mattikalli et al., “Universal lesion detection in CT scans using neural network ensembles”, In: SPIE Medical Imaging: Computer-Aided Diagnosis, 12033, (2022).



(a) VFNet (b) 快速 RCNN (c) 黄斑箱 (d) 视网膜网络

图 2: 列（a）至（d）显示了来自两位不同患者的 CT 体积切片的各种模型的输出。每对中的第一行代表在不平衡的数据集 D_U 上训练后的模型输出，而第二行则展示了使用按身体部位标签平衡的数据集 D_{BP} 进行训练的结果。绿色框：GT，黄色框：TP，红色框：FP。还显示了预测的类别和置信分数。第一对显示，在 D_U 上训练的模型未能正确识别和分类“Bone”病变（第一行），而在 D_{BP} 上训练的一个模型则可以做到（第二行）。特别是，VFNet 在 D_{BP} 上的训练以 97%的信心正确预测。第二对显示了使用 D_{BP} 时 VFNet 的 FP 较少，并且 FoveaBox 漏检（最后一行）。