将眼动数据融入胸部X光片深度学习方法：一项初步研究

Andr´e Lu´ıs INESC-ID / 里斯本大学 Instituto Superior T´ecnico，葡萄牙
andre.t.luis@tecnico.ulisboa.pt

Chihcheng Hsieh *信息系统学院 昆士兰科技大学，澳大利亚
chihcheng.hsieh@hdr.qut.edu.au

Isabel Blanco Nobre 影像学部 Grupo Lus´ıadas，里斯本，葡萄牙
isabel.blanco.nobre@lusiadas.pt

Sandra Costa Sousa 影像学部 Grupo Lus´ıadas，里斯本，葡萄牙
sandra.costa.sousa@lusiadas.pt

Anderson Maciel INESC-ID / 里斯本大学 Instituto Superior T´ecnico，葡萄牙
anderson.maciel@tecnico.ulisboa.pt

Joaquim Jorge INESC-ID / Instituto Superior T´écnico 里斯本大学，葡萄牙 jorgej@acm.org

Catarina Moreira INESC-ID / 信息系统学院 昆士兰科技大学
catarina.pintomoreira@qut.edu.au

arXiv:2302.02940v1 [cs.CV] 6 Feb 2023

摘要

本文提出了一种新颖的多模态深度学习架构，将医学图像与眼动追踪数据结合起来，用于胸部X光片的异常检测。我们的结果表明，将视线数据直接应用于深度学习架构，并不能提升胸部X光片异常检测的预测性能。这些结果与文献中的其他研究相符，并表明在将视线等人工生成数据应用于深度学习架构之前，还需要对其进行更深入的研究。

1 引言

诊断成像在诊断和治疗癌症、心脏病及损伤等多种疾病方面发挥着至关重要的作用。它使医疗专业人员能够及时准确地作出诊断，并确定最有效的治疗方案。然而，获得放射科医生服务仍是一项重大挑战，尤其是在医疗资源匮乏和偏远地区。全球放射科医生短缺，以及难以向偏远地区提供放射学服务，都加剧了这一问题。这可能导致诊断延迟、医疗费用增加以及患者预后变差。

深度学习（DL）已成为解决医疗保健服务不足和偏远地区诊断成像难题的有力方案。深度学习算法能够以很高的准确率从医学图像中提取并学习复杂特征，从而自动分析图像，识别人类放射科医生可能难以发现的模式和异常。这有助于减少对放射科医生的依赖，并改善医疗保健服务不足地区患者获得医疗服务的机会。然而，这些系统也容易受到偏差影响，且缺乏透明度，这可能导致放射科医生对其信任度和在医疗保健中的采用率下降。

虚拟现实（VR）技术有望改变放射科医生评估医学图像的方式，为解决放射科医师短缺和医疗保健可及性方面的挑战提供突破性方案 [29]。通过打造沉浸式虚拟放射科阅片室，虚拟现实使放射科医生能够远程工作，即使面临放射科医师短缺，也能提供准确的诊断。这在大流行病期间，这一点尤为重要：身体健康的放射科医生可能不得不留在家中，而放射科医师短缺可能导致诊断延误。此外，虚拟现实技术还可通过纳入眼动追踪等新的数据模态，为深度学习方法开辟新的研究路径。研究表明，虚拟现实中的眼动追踪能够估算认知负荷 [27, 31], 和疲劳 [28]，并检测真阳性／假阳性 [19]。通过提供有关放射科医生如何查看医学图像的宝贵信息 [18],，基于虚拟现实的眼动追踪有助于提高深度学习系统的诊断准确率 [21]。

本文介绍一种前沿多模态深度学习（DL）架构的初步研究结果。该架构整合医学图像和眼动追踪，将其作为一种新型模态，用于检测胸部X光片中的异常。本研究旨在评估这种方法提升深度学习系统诊断准确率的潜力。眼动追踪是测量视线的过程，可为诊断过程提供有价值的洞见，例如放射科医生关注的位置以及他们解读图像的方式。通过将这些数据纳入我们的深度学习架构，我们旨在减少对放射科医生的依赖，并改善医疗保健服务不足地区患者获得医疗服务的机会。此外，我们还旨在探究纳入眼动追踪能否使深度学习系统更能抵御偏差，并提高其诊断准确率。

2 相关工作

诊断成像是一个成熟的领域，长期以来的研究一直致力于提高医学影像的准确性和可及性。尽管越来越多的研究开始探讨深度学习在医学影像诊断中的应用 [34], ，但这些研究仍面临虚假相关性的检测问题，进而导致有偏且错误的诊断。这包括利用深度学习检测医学图像（如X光、计算机断层扫描和磁共振成像扫描）中异常的研究 [30]。这一问题的潜在解决方案是将多种数据模态纳入深度学习架构，以减少学习过程对医学图像的单一依赖。

先前的研究探讨了将眼动追踪数据与医学图像相结合，特别是用于训练目的 [15]。较新的研究则转向探讨眼动追踪在深度学习中的影响 [19]。EYEGAZE数据集 [14]和 REFLACX [2] 是此类数据集的两个例子。Eye-Gaze数据集包含一位放射科医生对 1,083 份正位胸部X光（CXR）影像的阅片记录，以及口述报告。

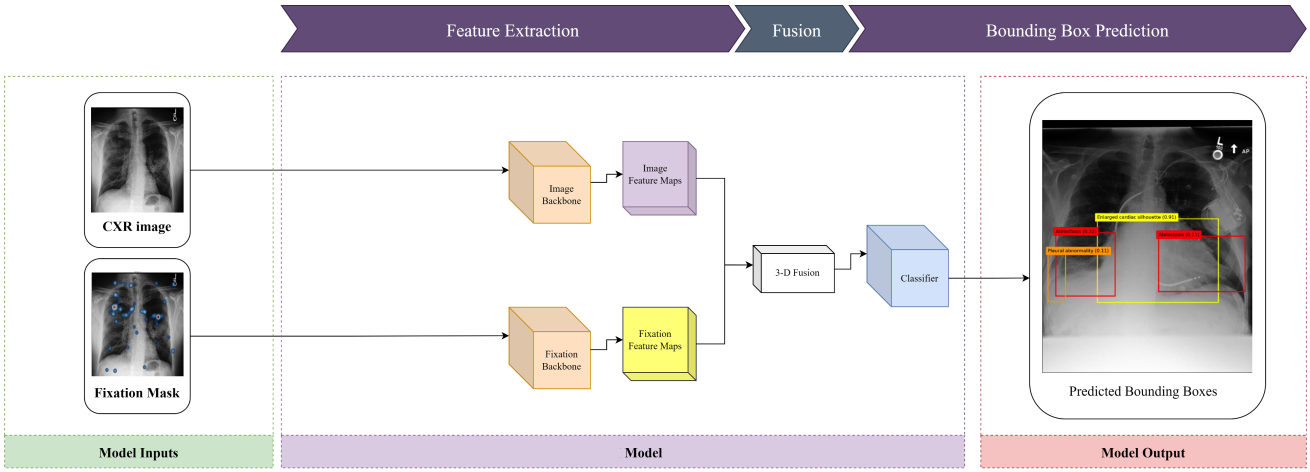


图1：所提出的多模态深度学习架构将注视图与胸部X光图像相结合，并对以下异常进行检测：五个类别：心脏轮廓增大、肺不张、胸膜异常、肺实变和肺水肿

所用图像取自 MIMIC数据库 [12]，图像级标签则通过自然语言处理技术从 MIMIC报告中提取。另一方面，REFLACX 包含 3,032 组同步的眼动追踪数据和带时间戳的报告转录文本，图像来自 MIMIC-CXR数据集。在此数据集中，五位放射科医生为胸部X光片标注了边界框及其标签。这些数据集为研究如何结合眼动追踪数据与医学图像以提高诊断准确率提供了宝贵资源。这两个面向医学图像的眼动追踪数据集向公众开放，推动了对新型多模态深度学习架构的研究，这些架构将放射科医生的眼动追踪数据与患者的医学图像信息相结合。

Karargyris等人 [14] 提出了两种深度学习架构，用于将眼动追踪数据纳入胸部X光（CXR）图像的图像级标签预测中。第一种方法将经过卷积神经网络（CNN）处理的CXR图像，与通过带自注意力机制 [9]的单层双向长短期记忆网络处理的时间注视热图进行拼接。结果显示，与仅使用CXR图像数据作为输入的基线模型相比，使用时间注视热图可使曲线下面积（AUC）提高5%。第二种方法使用静态注视热图，即将所有时间注视点聚合到单张图像中。在训练期间，模型对静态注视热图和图像级标签进行联合训练。在测试期间，当输入一张CXR图像时，模型会输出标签以及表示该病症最重要位置的热图分布。该方法取得了与基线模型相近的结果，同时增强了可解释性。

Wang等人也研究了深度学习中的注视数据。 [32], 他们收集了放射科医生的注视信息，并构建了一个注视引导的注意力网络，确保网络能够像放射科医生一样聚焦于病变区域，并根据标注的边界框和放射科医生的注视信息，输出疾病标签及注意力图。结果表明，在GA-net架构中使用放射科医生的注视作为监督信号，优于仅使用图像的最先进方法，例如 ResNet [33], 或视觉Transformer [5]。此外，与放射科医生手动标注边界框相比，收集注视数据速度更快，且分类准确率相近。其他研究表明，眼动追踪可能很有用，因为放射科医生最初的注视点通常与包含病灶的区域重合 [24]。

尽管将眼动追踪数据融入用于诊断成像的深度学习模型具有潜在优势，但文献中已指出若干挑战。一些研究指出，不应在深度学习中使用基于人类注视的显著性图，因为有证据表明，这些系统可能会利用背景上下文进行目标分类，从而产生偏差 [23,26]。此外，最早将眼动追踪数据用于胸部X线分析的研究之一 [3] 利用眼动追踪数据区分了三种诊断错误：（1）搜索错误，即未能发现目标；（2）识别错误，即目光注视到目标，但仍未能识别目标；（3）决策错误，即放射科医生未能报告检查结果。这些错误会污染眼动追踪数据的收集，并对深度学习模型的性能产生负面影响。研究报告称，这些错误的比例最高可达：搜索错误占30%、识别错误占25%、决策错误占45% [11, 16]。在设计和评估融入眼动追踪数据、旨在提高诊断准确率的深度学习模型时，必须考虑这些挑战。本研究旨在进一步探讨深度学习技术能否从医学诊断中的眼动追踪数据中获益。

3 一种采用注视掩膜的多模态深度学习架构

所提出的架构是对Mask R-CNN网络的扩展 [6]。Mask R-CNN是一种用于图像和视频帧中目标检测与实例分割的最先进算法。它使用区域提议网络（RPN）生成图像中可能包含物体的候选区域，然后通过一个独立网络对这些区域中的物体进行分类和分割。

Mask R-CNN 的关键创新是在区域提议网络和分类网络之上增加了一个全卷积网络分支，为图像中的每个对象实例生成二值掩膜。这样可以更精确地定位图像中的对象，因为它能够识别属于每个对象的具体像素。

Mask R-CNN 的架构由三个主要组成部分构成：骨干网络，用于从输入图像中提取特征图（通常采用预训练的卷积神经网络（CNN），例如 ResNet）；区域提议网络，根据特征图生成目标候选区域；以及检测网络，用于对候选区域中的对象进行分类和分割。

我们通过添加一个基于放射科医生眼动模式注视图的新骨干网络，扩展了该架构。由于 Mobilenet 在小型数据集上表现良好，我们将其用作骨干网络 [10]。图1展示了所提出的多模态模型架构。首先，采用逐元素方式将胸部X光图像和注视掩膜进行融合

求和。随后，区域提议网络输出一组矩形目标提议，池化层将感兴趣区域与输入对齐至单一特征图。最后，将这些结果展平并传递给分类器，分类器针对每个候选区域输出边界框坐标、类别标签和目标掩膜。

多模态方法的内部工作机制和输出与基线技术相似。关键区别在于输入。输入不再是单张胸部X光图像，而是图像和注视掩膜；该掩膜包含放射科医生阅片期间注视点的热力图。二者通过逐元素乘法进行融合，即将两种图像表征的特征逐元素相乘。

$$L = L_{classification} + L_{bbox} + L_{mask} \quad (1)$$

Mask R-CNN 架构中每个样本感兴趣区域的损失函数如公式1所示。分类损失和边界框损失与 Faster R-CNN [6],中的相同，这意味着分类损失对应于二元交叉熵（目标与非目标）。相比之下，回归损失（边界框）采用平滑L1范数。掩膜损失仅针对真实类别定义，且为平均二元交叉熵损失。在训练过程中，贯穿各个训练轮次监测这些损失；其下降趋势表明模型正在学习有意义的关系。

评估过程中，使用检测框面积交并比（IoBB）阈值来判定预测结果是否正确。只有当预测结果与真实边界框的交并比高于 IoBB 阈值时，才将其视为正确结果（真正例）。为考察模型的定位能力，我们还测试了模型在不同 IoBB 阈值下的表现。在医学领域，IoBB 比交并比（IoU）更为适用，因为真实边界框通常会设置得较大，以涵盖所有分散或较大的病灶。为了评估不同模型，我们采用了边界框交并比为 50% 时的绝对精确率和召回率。精确率表示病灶类别预测结果中属于病灶类别的比例，而召回率表示数据集中所有病灶样本中被预测为病灶类别的比例。这些架构的训练并非从零开始，而是基于在公开数据集上预训练的模型进行。本文测试了两种骨干网络：ResNet [8] 和 MobileNet [10], ，二者均在 Microsoft COCO数据集 [17]上进行了预训练。

4 数据集

REFLACX [2] 是一个包含3,032张胸部X光图像的报告和眼动追踪数据集，用于定位异常。它是首批公开可用的数据集之一，包含放射科医生进行胸部X光图像阅片时的眼动位置、瞳孔数据和口述记录。此外，五名放射科医生确定了异常边界框，这些边界框作为局部标签，可进一步用于目标定位。

向放射科医生展示的图像来自数据收集流程中从 MIMIC-CXR数据集 [12]随机抽取的样本。抽样后，排除了异常图像，原因包括图像缺失部分或发生翻转。

一次胸部X光片阅片是指一次单独的数据收集会话，在该会话中，一名放射科医生分析一名患者的一张胸部X光片。在会话期间，专业人员可以识别出零个、一个或多个异常位置椭圆。图2展示了用于收集数据的界面示例 [2] 以及放射科医生的眼动模式。

5 结果与讨论

我们比较了仅使用胸部X光片的 Mark-RCNN（基线模型）与所提出的多模态深度学习架构；后者结合了胸部X光片和注视图。表1展示了所得结果。

结果表明，仅使用图像的模型表现优于使用放射科医生注视掩膜的多模态方法。我们认为，结果不佳是因为注视掩膜由视线数据聚类而成，这意味着一张胸部X光图像中只包含少量注视点。这些注视点大多与真实标注无关，这表明仍需进一步研究如何从眼动数据中提取与胸部X光片病灶标注相关的感兴趣区域。此外，进一步分析表明，REFLACX 数据中包含大量噪声：记录并公开的眼动模式中，有些注视是与界面交互和评估胸部X光片时产生的。

单独考察模型在各类异常状况上的表现时，Mask-CNN 的性能优势十分明显。对于该模型，识别心影增大状况（AP@[BBOX=0.5] = 43%）并不困难。这是因为这类异常更局限于心脏周围。其余病灶的表现相近且较低，可能是因为它们大多同时出现在肺部区域，导致分类器难以区分。

尽管从性能角度来看，我们无法证明在多模态深度学习架构中使用眼动数据具有明显优势，但这项初步研究使我们得以为未来研究总结出若干见解：

1. 当前基于深度学习、使用 RE-FLACX 数据集的方法往往会直接使用原始眼动数据，而未考虑其人工生成过程的复杂性。这会产生噪声极大的数据流，且无法有效助力监督学习过程。

2. REFLACX采用的数据收集过程使用了传统的桌面式眼动追踪仪，这类设备容易受到噪声的严重干扰，而且在数据收集过程中不允许头部大幅度移动。这可能导致数据点缺失或记录的数据不准确。

3. 眼动追踪数据是一种以人为中心的数据。因此，在处理和分析此类数据时，必须考虑其生成过程中涉及的人为因素。这包括但不限于考察人类搜索模式。

6 结论与未来方向

眼动追踪数据已在医学影像领域得到广泛研究。近期，研究人员开始探索将眼动追踪数据融入深度学习方法的潜力，旨在开发新型深度学习架构，根据放射科医生的行为模式，实现更准确、精确的医学诊断。这些系统或可成为解决全球放射科医生短缺问题的一种可行方案。少数采用深度学习将眼动追踪数据与胸部X光图像相结合的研究结果不一：有些报告显示存在预测优势，而另一些则未发现这种优势。

在这项研究中，我们提出了一种基于 Mark-RCNN 的多模态深度学习架构，将放射科医生的注视图与胸部X光图像相结合，以探究这种差异。结果显示，纳入注视图并未带来性能优势。

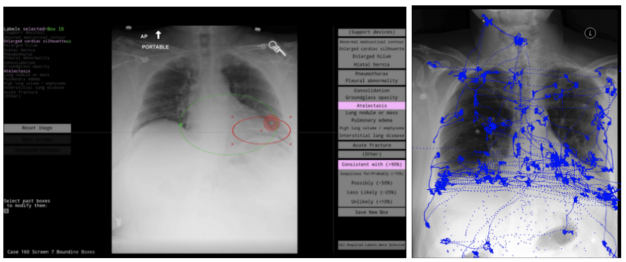


图2：REFLACX 数据采集界面示例（左）以及记录的放射科医生眼动数据。

	掩码卷积神经网络（仅图像）		掩码卷积神经网络（图像 + 注视点）	
异常	平均精度@[IoBB=0.50]	平均召回率@[IoBB=0.50]	平均精度@[IoBB=0.50]	平均召回率@[IoBB=0.50]
心影增大	0.429326	0.810000	0.436229	0.760000
肺不张	0.125410	0.529412	0.092772	0.192513
胸膜异常	0.043422	0.226519	0.052553	0.165746
肺实变	0.113867	0.308642	0.010061	0.086420
肺水肿	0.030728	0.410959	0.001856	0.027397
平均值	0.148551	0.457106	0.246415	0.144261

表1：使用仅包含胸部X光片的基线模型 Mask R-CNN，以及将胸部X光片图像与从眼动模式中提取的放射科医生注视图相结合的多模态深度学习架构所获得的结果。这些结果支持了相关文献中的观点，即将原始眼动追踪数据应用于深度学习方法并不会带来预测优势。

未来，我们计划扩展这项研究，分析放射科医生阅读胸部X光片时的瞳孔扩张情况。文献表明，瞳孔扩张与认知负荷之间存在很强的相关性 [1], , 因此，与注视图相比，瞳孔扩张可能是衡量其与胸部X光片异常区域相关性的更好指标。

通过本研究，我们还证实，用于收集数据的桌面式眼动追踪器极易受到噪声和头部运动的影响。未来，我们还计划使用虚拟现实眼镜复现 REFLACX 的数据收集过程。在虚拟现实环境中，光线和对比度条件可以轻松控制，放射科医生的头部运动不会受到限制，而且眼动数据收集过程中的噪声更少。

致谢

本研究部分得到了联合国教科文组织人工智能与扩展现实教席以及葡萄牙科学技术基金会~（*FCT*）的支持，资助项目编号为 2022.09212.PTDC 和 UIDB/50021/2020。

参考文献

[1] R. Bednarik等人. 瞳孔大小可作为显微外科训练任务中视觉运动负荷和专业水平的指标. 收录于*2018*年*ACM*眼动追踪研究与应用研讨会论文集，ETRA '18，美国纽约州纽约市，2018年.[2] R. Bigolin Lanfredi等人. Reflacx：一个包含胸部X光片异常定位报告和眼动追踪数据的数据集. *Scientific Data*，9(1):1–15，2022年. [3] D. Carmody、C. Nodine和H. Kundel. 通过比较性视觉扫描和不进行比较性视觉扫描来寻找肺结节. 感知与心理物理学，29:594–598，1981年11月. [4] N. Correia等人. Instory：一个用于实体空间中的移动信息访问、讲故事和游戏活动的系统. 收录于*2005*年*ACM SIGCHI*国际娱乐技术进展会议论文集，ACE '05，第102–109页. [5] A. Dosovitskiy等人. 一张图像胜过16×16个词：用于大规模图像识别的Transformer. *CoRR*，abs/2010.11929，2020年. [6] R. B. Girshick. Fast R-CNN. *CoRR*，abs/1504.08083，2015年. [7] D. J. Gonc̦alves和J. A. Jorge. 个人文档空间的实证研究. 收录于J. A. Jorge、N. Jardim Nunes和J. Falc˜ao e Cunha编，交互式系统：设计、规范与验证，第46–60页，柏林、海德堡，2003年. Springer Berlin Heidelberg. [8] K. He等人. 用于图像识别的深度残差学习. *CoRR*，abs/1512.03385，2015年. [9] S. Hochreiter和J. Schmidhuber. 长短期记忆. 神经计算，9:1735–80，1997年12月. [10] A. G. Howard等人. MobileNets：面向移动视觉应用的高效卷积神经网络. *CoRR*，abs/1704.04861，2017年. [11] C. H. Hu等人. 搜索骨折：与肺结节搜索的比较. *Academic Radiology*，1(1):25–32，1994年. [12] A. E. W. Johnson等人. MIMIC-CXR：大型公开标注胸部放射影像数据库. *CoRR*，abs/1901.07042，2019年. [13] J. Jorge和F. Samavati. 基于草图的界面与建模. Springer Science & Business Media，2011年.

[14] A. Karargyris等人. 创建并验证用于人工智能开发的胸部X光数据集，其中包含眼动追踪数据和报告口述内容. Scientific Data，8(1):1–18，2021年. [15] C. Kosel、D. Holzberger和T. Seidel. 识别专家和新手的视觉扫描路径模式，并探究其与评估学习相关学生特征之间的关系. Frontiers in Education，5，2021年.[16] E. A. Krupinski. 放射科医生搜索乳房X线影像时的视觉扫描模式. Academic Radiology，3 2:137–44，1996年. [17] T. Lin等人. Microsoft COCO：情境中的常见物体. CoRR，abs/1405.0312，2014年. [18] L. L´evˆeque等人. 研究现状：医学影像领域的眼动追踪研究. IEEE Access，6:37023–37034，2018年. [19] S. Mall、P. Brennan和C. Mello-Thoms. 乳房X线影像中的注视区域与未注视区域：对视觉搜索行为的高阶统计分析. Academic Radiology，2017年. [20] C. Moreira 、 B. Martins 和 P. Calado. 使用排序聚合在学术数字图书馆中搜索专家，2015年. [21] C. Moreira等人. 通过眼动追踪和扩展现实改进X光诊断. 收录于2022年IEEE虚拟现实与三维用户界面会议摘要与研讨会（VRW），第450–453页，2022年. [22] C. Moreira和A. Wichert. 学习对学术专家进行排序. 硕士论文，Instituto Superior Técnico，2011年. 荣誉与优异奖（Summa cum laude）. [23] W. Nie、Y. Zhang和A. Patel. 对基于反向传播的可视化中令人困惑行为的理论解释. CoRR，abs/1805.07039，2018年. [24] C. F. Nodine和H. L. Kundel. 利用眼动研究视觉搜索并改进肿瘤检测. RadioGraphics，7(6):1241–1250，1987年. [25] J. a. P. Pereira等人. 用于模糊书法式交互的级联识别器. 收录于首届Eurographics基于草图的界面与建模会议论文集，SBM'04，第63–72页，德国戈斯拉尔，2004年. Eurographics Association. [26] Z. Qi、S. Khorram和F. Li. 通过使用积分梯度进行优化来可视化深度网络. CoRR，abs/1905.00954，2019年. [27] K. Rayner和E. M. Reingold. 阅读过程中注视时长受直接认知控制的证据. Current Opinion in Behavioral Sciences，1:107–112，2015年. 认知控制. [28] E. D. Reichle等人. 使用E-Z Reader模拟非阅读任务中的眼动：理解眼动—心智联系的统一框架. Psychological Review，119(1):155–185，2012年. [29] M. Sousa等人. Vrrrroom：阅片室中的放射科医生虚拟现实体验. 收录于2017年CHI计算机系统中的人为因素会议论文集，第4057–4062页. ACM，2017年. [30] W. Tan等人. 大数据时代的多模态医学图像融合算法. Neural Computing and Applications，2020年. [31] O. Tolvanen等人. 外科手术中工作负荷的眼动追踪指标：系统综述. Journal of Investigative Surgery，35:1–10，2022年1月. [32] S. Wang 等人 . 跟随我的视线：利用注视监督计算机辅助诊断 . IEEE Transactions on Medical Imaging，2022年. [33] S. Woo等人. CBAM：卷积块注意力模块 . CoRR，abs/1807.06521，2018 年 . [34] E. C̦allı等人. 用于胸部X光影像分析的深度学习：综述. Medical Image Analysis，72:102125，2021年.