重新思考计算机辅助下的肺结核诊断*

Yun Liu[1†] Yu-Huan Wu[1†] Yunfeng Ban[2] Huifang Wang[2] Ming-Ming Cheng[1✉]
[1]Nankai University [2]InferVision
http://mmcheng.net/tb/

Abstract

肺结核是一种严重的传染病，是世界范围内对人类健康的主要威胁之一，每年会造成数百万人死亡。虽然早期肺结核诊断和治疗可以大大提高生存机会，但肺结核诊断对大家来说，特别对发展中国家来说仍然是一个重大挑战。计算机辅助下的肺结核诊断 *(Computer-aided Tuberculosis Diagnosis, CTD)* 是肺结核诊断的一种很有前景的选择，但训练数据的缺乏阻碍了 *CTD* 的发展。为了解决这个问题，本文建立了一个大规模的肺结核数据集，名为 *TBX11K*。相对现有最大的公共肺结核数据集只有 *662* 张 *X* 光片图像具有相应的图像类别标签的标注，本文提出的数据集包含 *11200* 张 *X* 光片图像，并在每张 *X* 光片中对肺结核病灶区域使用边界框进行了标注。在足够多的数据支撑下，该数据集能够支持深度检测器的训练，从而得到高质量的计算机辅助肺结核诊断结果。本文还对现有的目标检测器进行了改进，使其适应于同时进行图像分类和肺结核区域检测。本文在 *TBX11K* 数据集上对这些改进的检测器进行了训练和评估，并作为未来研究的基准。

1. 引言

肺结核 (Tuberculosis, TB) 是第二大致死性传染病 (仅次于艾滋病毒)，也是全球主要的健康威胁之一 [34, 35]。每年大约有八百万至一千万的新肺结核患者，大约有两百万至三百万人死于肺结核 [35]。肺结核是由结核分枝杆菌引起的，结核分枝杆菌可通过打喷嚏、咳嗽或其他排出传染性细菌的方式传播。因此，肺结核通常通过呼吸道在肺部发生。而发展中国家中艾滋病毒患者和营养不良者等免疫功能低下者的机会性感染更是加剧了这一问题。

如果不进行治疗，肺结核的死亡率非常高，但是在早期诊断出肺结核并使用抗生素进行治疗大大提高了生存的机会 [6,17,19]。肺结核的早期诊断也有助于控制感染的传播 [6]。具有多种耐药性的肺结核的增加也导致本文迫切需要一种及时、准确的肺结核诊断方法，以跟踪临床治疗的过程 [11]。不幸的是，肺结核诊断仍然是一个重大挑战 [1, 2, 5, 6, 17, 19, 32]。肺结核诊断的金标准是通过痰液的显微镜检查和结核菌培养，并检测是否存在结核分枝杆菌来界定 [1, 2] 。但是，培养结核分枝杆菌需要生物安全三级实验室 (BSL-3)，而这个过程通常需要几个月的时间 [1, 2, 19]。更糟糕的是，许多发展中国家和资源有限的社区的医院无法提供这样的条件。

另一方面，X 光片检查是目前医学图像检查中最常见的、数据密集的筛查方法，也是肺结核筛查最常用的方法之一。X 光片的早期筛查对肺结核的早期发现、治疗和预防具有重要意义 [5,19,21,36,41]。然而，放射科医生对 X 光片的检查结果经常出错 [21, 36]，因为人眼通常很难从 X 光片中区分出肺结核区域，人眼也对 X 光片的很多细节不够敏感。在本文的人类准确度研究中，与金标准相比，来自顶级医院的经验丰富的放射科医生的准确率只有 *68.7%* 。

*本文为 [31] 的翻译版。
†Joint first authors.

动机和贡献 得益于深度学习，特别是卷积神经网络 (CNNs) [12, 15, 38] 的强大表现能力，深度学习在人脸识别 [39] 、图像分类 [14]、目标检测 [13] 和边缘检测 [29] 等领域都做的比人类好。深度学习可以捕捉细节 [16, 28, 29] , 而且从来不会像人一样感到疲倦。将深度学习应用于计算机辅助下的 X 光片肺结核诊断/筛查是一个很自然的想法。然而，深度学习的训练总是需要大量的数据，而收集大规模的肺结核数据是困难的，因为它们非常昂贵和私密。

由于缺乏公可开获得的 X 光片，**计算机辅助下的肺结核诊断** (**CTD**) 无法成功地利用深度学习来提高性能。例如，目前最大的用于肺结核诊断的公共 X 光片数据集是 [18] 中提出的 Shenzhen 数据集。Shenzhen 数据集包含 662 张 X 光片图像，包括 336 张结核表现的 X 光片和 326 张正常的 X 光片，即只有两类图像类别的标签。仅仅使用这几百张图片不足以训练深度卷积神经网络 (CNNs)。因此，许多先进的 CTD 方法只采用手工标注的特征 [5, 19, 20] 或预先训练好的 CNNs 作为特征提取器，而没有进行微调 [32]，忽略了深度 CNNs 强大的自动特征学习能力。

为了实际部署 CTD 系统来帮助世界各地的肺结核患者，本文必须首先解决数据不足的问题。在本文中，本文通过与各大医院的长期合作，为研究者社区提供 **肺结核 X 光片** (**TBX11K**) 数据集，这个新的数据集在以下方面优于先前的 CTD 数据集 i) 与之前的只包含几十或几百张 X 光片图像的数据集 [6, 18] 相比，TBX11K 有 11,200 张图像，约为最大现有的数据集即深圳数据集 [18] 的 17 倍，所以 TBX11K 令训练很深的卷积神经网络成为可能; ii) TBX11K 不像以前的数据集只有图像类别标签的标注，而是使用边界框对肺结核区域进行标注，使未来的 CTD 方法不仅可以识别是否存在肺结核，还可以检测对应的肺结核区域，从而帮助放射科医生进行更准确的诊断; iii) TBX11K 包括四种类别，分别为健康、活动性肺结核 (Active TB)、陈旧性肺结核 (Latent TB) 和患病但非肺结核，而不是其他数据集中对肺结核或非肺结核的二分类，因此未来的 CTD 系统可以适应更复杂的现实场景，并为人们提供更详细的疾病分析。TBX11K 中的每一张 X 光片图像都**使用金标准** (即微生物学诊断法) 进行测试，然后由各大医院经验丰富的放射科医生进行标注。TBX11K 数据集已经被数据提供商去敏感化，并被相关机构豁免，可以公开使用，以促进未来的 CTD 研究。

数据集名称	年份	类别数	标注类型	Sample
MC [18]	2014	2	图像类别	138
Shenzhen [18]	2014	2	图像类别	662
DA [6]	2014	2	图像类别	156
DB [6]	2014	2	图像类别	150
TBX11K	2020	4	边界框	11200

表 1. 目前公开的肺结核数据集概要。

此外，本文对现有的 SSD [27]，RetinaNet [25]，Faster R-CNN [37]，和 FCOS [40] 等目标检测方法进行了改良，使它们同时进行图像分类和 TB 区域检测。具体而言，本文在这些检测器上引入了一个分类分支，并提出了另一种训练策略。这些改良后的方法可以被视为未来 CTD 研究的基准方法。在本文提出的 TBX11K 数据集上，本文还把图像分类和目标检测的度量标准引入到了肺结核 (TB) 诊断的评价中，从而根据这些度量标准来评估本文构建的基准方法，以构建肺结核诊断初始的基准评价。

总之，本文的贡献有两方面:

- 本文构建了一个大规模 CTD 数据集，它比现有肺结核数据集大得多、标注更精细、更真实，从而能够训练深度 CNNs。
- 本文通过以下方式建立了 CTD 的基准: i) 改良已有的用于 CTD 的目标检测器; ii) 引入分类和检测指标以用于 CTD 的评价，这预计将为未来的 CTD 研究奠定一个良好的开端。

2. 相关工作

2.1. 肺结核数据集

由于肺结核数据是非常私密的，而且很难用金标准诊断肺结核，因此公开获得的肺结核数据集非

常有限。本文在 Tab. 1提供了公开的肺结核数据集的概要。其中 Jaeger 等人 [18] 提出了两种用于肺结核诊断的胸片数据集。蒙哥马利县胸片集 (MC) [18] 是与美国马里兰州蒙哥马利县卫生和公共服务部合作收集的。MC 数据集包括 138 张 X 光片照片，其中 80 张为健康病例，58 张为有结核表现的病例。深圳胸透片集 (Shenzhen) [18] 是通过与中国广东省深圳市第三人民医院、广东医科大学合作采集的。深圳数据集由 326 例正常案例和 336 例有结核表现的案例，X 光片图像共 662 张。Chauhan 等人 [6] 提出了两个数据集，名为 DA 和 DB，它们是从新德里国家肺结核和呼吸疾病研究所的两台不同的 X 光机器中获得的。DA 由训练集 (52 张 TB 和 52 张非 TB X 光片) 和独立测试集 (26 张 TB 和 26 张非肺结核 X 光片) 组成。DB 包含 100 张训练 X 光片 (50 张 TB 和 50 张非 TB) 和 50 张测试的 X 光片 (25 张肺结核和 25 张非肺结核)。注意，这四个数据集都只使用了图像类别标签标注了图像的二分类。

由于这些数据集太小，人们无法利用这些数据训练深度神经网络。因此，尽管 CNNs 在计算机视觉领域取得了许多成功，但 CTD 的最新研究一直受到阻碍。另一方面，现有的数据集只有图像类别标签级别的标注，因此本文不能用以前的数据训练肺结核检测器。为了帮助放射科医生做出准确的判断，本文需要检测肺结核区域，而不仅仅是图像级别的分类。因此，由于肺结核数据的缺乏，深度学习无法为实际的 CTD 系统带来成功，而这些系统有可能每年拯救数百万肺结核患者。

2.2. 计算机辅助下的肺结核诊断

由于数据的缺乏，传统的 CTD 方法无法训练深度 CNNs。大多数传统方法主要使用手工标注的特征和训练二元的分类器。Jaeger 等人 [19] 首先使用图割 (graph cut) 方法 [4] 对肺区域进行了分割。然后，他们从这个肺区域利用手工设计的方法提取了结构和形状特征。最后，他们使用二元分类器，即支持向量机，将 X 光片分为正常和非正常类别。Candemir 等人 [5] 将基于图像检索的患者特异适应性肺模型用于非刚性的、注册驱动且鲁棒的肺分割方法，有利于传统的肺特征提取 [19]。Chauhan 等人 [6] 实现了一个 MATLAB 工具箱 TB-Xpredict，该工具箱采用 Gist [33] 和 PHOG [3] 特征，不需要分割就可以区分肺结核和非肺结核 X 光片 [8,30]。Karargyris 等人 [20] 提取了形状特征来描述肺的整体几何特征，并提取纹理特征来表示图像特征。

Lopes 等人 [32] 没有使用手工标注的特征，而是采用 ImageNet [10] 上预训练的固定 CNNs 作为特征提取器，从而计算 X 光片图像的深度特征。然后，他们训练了一个 SVM 对这些深度特征进行分类。Hwang 等人 [17] 使用私有数据集训练 AlexNet [22] 来进行二分类 (肺结核和非肺结核)。其他私有数据集也在 [23] 中用于图像分类网络。然而，本文提出的数据集即 TBX11K ，将会对外公开，以促进这一领域的研究。

3. 肺结核 X 光片 (TBX11K) 数据集

3.1. 数据采集与标注

对于数据的采集和标注，本文有三个步骤:1) 建立分类类别，2)X 光片采集，3) 专业的数据标注，具体介绍如下。

3.1.1 分类类别建立

现有的肺结核数据集只包含两类: 肺结核和非肺结核，其中非肺结核指的是健康案例。在实际应用中，胸部 X 光异常，如结核、肺不张、心脏肿大、积液、浸润、肿块、结节等，具有相似的异常形态 (如病灶模糊、不规则)，它们与健康的 X 光明显不同，健康的 X 光几乎具有同样清晰的样子。因此，仅将健康 X 光片作为阴性类别，在模型预测有许多患者但非肺结核患者的临床场景时，会有较大的偏差，也会导致大量的假阳性。为了将 CTD 推广到实际应用中，本文在本文的数据集中考虑了一个新的类别，即患病但非肺结核。此外，除了肺结核的识别外，区分活动性肺结核和陈旧性肺结核也非常重要。活动性肺结核是由结核分枝杆菌感染或陈旧性结核病的重新激活引起的，而陈旧性肺结核患者既不生病也不具有传染性。区分活动性肺结核和陈旧性肺结核

可以帮助医生为患者提供适当的治疗。考虑到这一点，本文在数据集中将肺结核分为两类: 活动性肺结核和陈旧性肺结核。根据上述分析，本文在提出的 TBX11K 数据集中包括四个类别: 健康、患病但非肺结核、活动性肺结核和陈旧性肺结核。

3.1.2 X 光片采集

肺结核 X 光片的收集面临两个难题: i) 胸片, 特别是肺结核 X 光片的隐私性较高，泄露这些数据会使人们面临违法风险，个人几乎不可能获取原始数据; ii) 虽然全世界有数百万肺结核患者，但由于结核分枝杆菌的检查过程复杂而漫长 (即几个月 [1,2])), 能够通过金标准进行明确检测的肺结核 X 光片很少。为了克服这些困难，本文作者与顶级医院合作收集了相关数据。本文得到的 TBX11K 数据集包含 11200 张 X 光片，包括 5000 例健康案例、5000 例患病但非肺结核病例和 1200 例有肺结核表现的病例。在这里，每一张 X 光片都对应不同的人。这 1200 张肺结核 X 光片由 924 例活动性肺结核病例、212 例陈旧性肺结核病例、54 例同时包含活动型结核和陈旧性结核的病例以及 10 例在当今医疗条件下无法识别其结核类型的不确定肺结核病例组成。本文收集了 5000 例患病但非肺结核的病例，以便在临床情况下尽可能多地覆盖 X 光片检查疾病类型。所有 X 光片的分辨率都在 3000×3000 左右。本文的数据集还包括每张 X 光片相应的性别和年龄，为肺结核诊断提供更全面的临床信息。这些数据已经被数据提供商去敏感化和相关的政府机构豁免，因此可以合法地公开这些数据集。

3.1.3 专业的数据标注

本文数据集中的每一张 X 光片图像都已经使用金标准进行了明确的测试，但是金标准只能提供图像类别标签的标注。例如，如果一个病人的痰液有结核表现，本文可以知道相应的 X 光片属于肺结核的哪种类别，但本文不知道肺结核在 X 光片中的确切位置和区域。另一方面，检测肺结核病灶区域对帮助放射科医生做出最终决定至关重要。仅通过图像类别层面的预测，人眼仍然很难找到肺结核区域, 这一点可以通过临床检查中放射科医生的低准确性得到证明，如 Sec. 3.3所示。如果 CTD 系统能同时提供图像分类和肺结核定位结果，那么放射科医师通过观察发现的肺结核区域，可以更准确、更有效地做出决定。

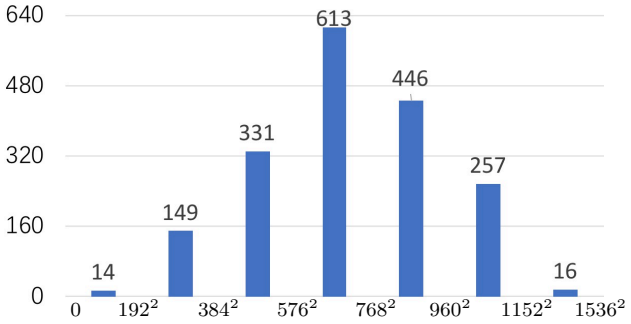

图 1. 肺结核边界框的区域分布。每个柱的左右值定义其对应的区域范围，每个柱的高度表示该区域内的肺结核边界框的数量。注意 X 光片的分辨率约为 3000×3000。

为了实现上述目标，本文为 TBX11K 数据集中的肺结核 X 光片提供了边界框标注。据本文所知, 这是肺结核检测的第一个具有边界框标注的数据集。所有的边界框都是由来自顶级医院的经验丰富的放射科医师进行标注的。具体来说，每一张肺结核 X 光片首先由具有 5-10 年肺结核诊断经验的放射科医生进行标记。然后，它的边界框标注会由另一位在肺结核诊断方面有超过 10 年经验的放射科医生进一步检查。它们不仅标记了肺结核区域的边界框，而且还会识别每个边界框的肺结核类型 (活动性肺结核或陈旧性肺结核)。本文通过对标记的肺结核类型进行了双重检查，确保它们与金标准的图像类别标签相一致。如果发生不匹配，此 X 光片将被放入未标记数据中重新标注，而标注者不知道之前哪个 X 光片曾经被标注错误。如果一个 X 光片被错误地标记了两次，本文会告诉标注者这个 X 光片的金标准检查结果，并请他们讨论如何重新标注。这种双重检查的过程使得这些标注好的边界框在肺结核区域检测中非常可靠。此外，非肺结核 X 光片仅具有由金标准得到的图像类别标签的标注。本文在 Fig. 3 中展示了提出的 TBX11K 数据集的一些例子。本文在 Fig. 1. 中显示了肺结核边界框的区域分布。大多数

	类别	训练集	验证集	测试集	总计
非肺结核	健康	3000	800	1200	5000
	患病且非肺结核	3000	800	1200	5000
肺结核	活动性肺结核	473	157	294	924
	陈旧性肺结核	104	36	72	212
	两种皆有	23	7	24	54
	不确定性肺结核	0	0	10	10
总计		6600	1800	2800	11200

表 2. TBX11K 数据集的划分情况。“活动性肺结核和陈旧性肺结核”是指同时含有活动性肺结核和陈旧性肺结核的 X 光片。“活动性肺结核”和“陈旧性肺结核”分别是指仅包含活动性肺结核和潜伏肺结核的 X 光片。“不确定性肺结核”是指在当今的医疗条件下无法识别其结核类型的结核 X 光片。

肺结核边界框的面积在 $(384^2, 960^2]$ 范围内。

3.2. 数据集子集划分

本文将数据分成三个子集，分别用于训练、验证和测试。本文的划分细节汇总在 Tab. 2中。为了更具代表性，本文考虑了四种不同的肺结核病例: i) 仅出现活动性肺结核; ii) 仅出现陈旧性肺结核; iii) 在 X 光片中同时出现活动性肺结核和陈旧性肺结核; iv) 无法识别肺结核类型的不确定性肺结核。对于各种肺结核病例，用于训练、验证和测试的肺结核 X 光片的数量比为 3 : 1 : 2。请注意，不确定性的肺结核 X 光片全部被放入到了测试集，研究人员可以使用这 10 张不确定性的 X 光片对类别不确定的肺结核检测进行评估。与科学实验设置一致，本文建议研究人员在训练集上训练他们的模型，并且在调整超参数时使用验证集进行评估。一旦他们的模型固定，他们就可以使用训练集和验证集进行再次训练，然后报告他们的模型在测试集上的性能。

3.3. 人类准确度研究

放射科医师的人体研究对本文了解 CTD 在临床肺结核诊断中的作用至关重要。本文首先从本文构建的 TBX11K 数据集的测试集中随机选择 400 张 X 光片，包括 140 张健康 X 光片、140 张患病且非肺结核 X 光片和 120 张肺结核 X 光片。120 张肺结核 X 光片包括 63 例活动性肺结核，41 例陈旧性肺结核，15 例活动性、陈旧性并存的肺结核和 1 例不确定性肺结核。然后，本文邀请了一位有 10 年以上工作经验的大医院的放射科医生，用图像级标签对这些 X 光片标注，标签从四个类别健康、患病但非肺结核、活性肺结核和潜伏肺结核中选定。如果活动性肺结核和陈旧性肺结核同时出现，该医生就指定这张 X 光片同时具有活动性肺结核和潜在性肺结核。请注意，这位放射学家与标记本文数据集的放射学家是不同的。

与金标准产生的正确结果相比，放射学家只能达到 68.7% 的准确性。若不对活动型和陈旧性肺结核加以区分, 准确率为 84.8% , 但肺结核类型的识别对临床治疗很重要。这样的较差表现是肺结核诊断、治疗和预防的主要挑战之一。与自然彩色图像不同的是，胸部 X 光片是灰度图像，通常有失真的和模糊的情况，这给识别带来很大的困难。然而，用金标准诊断肺结核需要几个月的时间 [1, 2] ，而且世界上许多地方没有这种条件。肺结核诊断方面的挑战是肺结核成为全球第二大传染病 (仅次于艾滋病毒) 的主要原因之一。在接下来的研究中，本文将展示在本文提出的 TBX11K 数据集上训练的深度学习 CTD 方法在诊断方面可以显著地优于经验丰富的放射科医生。

3.4. 潜在的研究主题

利用本文提出的 TBX11K 数据集，本文可以对 X 光片图像分类和肺结核区域检测进行研究。由于存在许多健康和患病但非肺结核的数据，本文的测试集可以模拟临床数据分布来评估 CTD 系统。本文认为开发同时进行 X 光片图像分类和肺结核区域检测的系统将是一个具有挑战性和有趣的研究课题。部署这样的系统来帮助放射科医生诊断肺结核是很方便的。

除了同时检测和分类，本文的数据集的另一个挑战是不同类别的数据分布是不平衡的。然而，这种数据不平衡与实际的临床情况是一致的。直觉上，当一个人去医院做胸部检查时，他很可能会感到不舒服，所以得病的概率比平常要高，但肺结核只是众多胸部疾病中的一种。在提出的 TBX11K 数据集中，本文假设只有 44.6% 的检查者是健康的，44.6%

是生病但非肺结核，只有 10.7% 的检查者感染了肺结核。陈旧性肺结核可通过两种途径引起:i) 接触活动性肺结核和 ii) 治疗后从活动性肺结核转化。大多数陈旧性肺结核是由第一种方式引起的。在医院的陈旧性肺结核患者通常是上述第二种，因为陈旧性肺结核患者既不生病，也不具有传染性，不太可能去医院检查。因此，本文的数据集中活动性肺结核病例远远多于陈旧性肺结核病例。因此，未来的 CTD 方法应该设计克服实际中的数据不平衡问题，例如，如何在类别不平衡的 TBX11K 训练集上训练模型。

4. 实验设置

在本节中，本文首先为同时进行 X 光片图像分类和 TB 区域检测的问题建立一些基准方法。然后，本文对评价指标进行了详细阐述。

4.1. 基准方法

现有的对象检测器不考虑背景图像。更具体地说，他们通常会忽略这些没有边界框框选对象的图像 [9, 25, 27, 37, 40, 42]。直接将现有的目标检测方法应用到 CTD 任务中，会导致检测出许多假阳性，因为实际中非肺结核的 X 光片的数量很大。为了解决这个问题，本文提出同时进行 X 光片图像分类与肺结核区域检测的方案，这样图像分类结果可以过滤出假阳性的检测结果。

本文对 SSD [27], RetinaNet [25], Faster R-CNN [37], and FCOS [40] 等著名的目标检测方法进行了改良，用于同时进行 X 光片分类和 TB 区域检测。图像分类的分支学习将 X 光片分为三**个类别**，即使用 *Softmax* 函数将图像分为健康、患病且非肺结核和肺结核三类。肺结核检测分支学习检测**两种类别肺结核的边界框**，即活动性肺结核和陈旧性肺结核。在临床诊断中，图像分类结果可以帮助放射学科医生判断肺结核是否出现在 X 光中。然后，肺结核检测结果为放射科医生提供肺结核病灶区域，帮助放射科医生做出最终决定。根据以上定义，本文在现有的目标检测器主干网络的最后的卷积层之后，增加一个图像分类分支，即在 VGG16 [38] 的 *conv5_3* 或 ResNet-50 [15] 的 *res5c* 之后。对于分类分支，本文使用 5 个连续的卷积层，每个层有 512 个输出通道和 ReLU 激活。第一卷积层中仅 SSD 方法的步长为 2，其他方法的步长为 1。在第三个卷积层之后连接一个最大池化层，步长为 2。在这些卷积之后，本文使用一个全局平均池化层和一个有 3 个输出神经元的全连接层来分为 3 类。由于本文的重点是为所提出的数据集的分析提供一些可行的基准，所以本文没有进行复杂的参数调整。

本文采用了一个二阶段训练的策略来训练这类网络。首先，本文省略图像分类分支，并使用默认设置训练目标检测器。然后，本文固定主干网络和目标检测分支，只训练图像分类分支从而使目标检测特征能够适应图像分类。第一阶段的训练只使用 TBX11K 训练集和验证集中的肺结核的 X 光片。第二阶段的训练不仅使用了所有训练集和验证集中的 X 光片，也随机使用了一半 MC [18] 和 Shenzhen [18] 数据集的训练集以及 DA [6] 和 DB [6] 的训练集。将 MC [18] 和 Shenzhen [18] 数据集的另一半以及 TBX11K, DA [6] 和 DB [6] 数据集的测试集用于评价图像分类的性能。肺结核区域检测的评估有两种模式: i) 使用全部 (即肺结核和非肺结核)TBX11K 测试集 X 光片和 ii) 仅使用 TBX11K 测试集中的 TB 的 X 光片。

所有实验都基于开源的 mmdetection 工具箱 [7] 和 4 个 RTX 2080Ti 计算卡。批大小是 16。第一阶段的训练分别迭代 38400 次 (在有 ImageNet 预训练 [10] 的前提) 或 76800 次 (从头开始训练)。初始学习率为 0.005, 但 SSD [27] 使用 0.0005。ImageNet 预训练时经过 25600 次和 32000 次迭代，或者从头开始训练时经过 51200 次和 64000 次迭代后，学习率除以 10。本文在第二阶段训练 24 个数据循环 (epoches), $1e$-3 的初始学习在第 12 个数据循环和第 18 个数据循环后被除 10。输入网络时，X 光片的图像大小调整为 512×512。

4.2. 评价指标

在本节中，本文将介绍评价 CTD 任务的度量标准。对于 X 光片的分类，CTD 的目标是将每一幅 X 光片分为三类，通过六个指标进行评估:

- 准确率，即 X 光片被正确分类为三类之一的百分比;
- 曲线下面积 (AUC)，用于计算受试者工作特征 (Receiver Operating Characteristic, ROC) 曲线下面积，曲线由肺结核类真实阳性率与假阳性率的构成;
- 敏感率，测量正确识别为肺结核的肺结核病例百分比，即肺结核类别的召回情况;
- 特异率，衡量被正确识别为非肺结核的非肺结核病例的百分比，即非肺结核类别的召回情况，其中非肺结核包括健康和疾病但非肺结核类别;
- 平均准确率 (AP)，计算每个类的精度并在所有类中取平均值;
- 平均召回率 (AR)，计算每个类的召回率和所有类的平均召回率。

对于肺结核检测的评价，本文采用 Microsoft COCO 数据集 [26] 提出的边界框平均精度 (AP^{bb})。默认的 AP^{bb} 是指 AP^{bb} 在不同边界框交并比阈值 [0.5 : 0.05 : 0.95] 下的平均精度。AP^{bb}_{50} 表示 AP^{bb} 的边界框交并比阈值为 0.5 下的平均精度。为了便于对每种肺结核类型的检测进行观察，本文分别报告活动性肺结核和陈旧性肺结核的评价结果。这里，不确定性肺结核的 X 光片被忽略了。本文还报告类别无关的肺结核检测结果，即忽略肺结核类别，以描述所有肺结核区域的检测情况。这里，不确定性肺结核的 X 光片也被包括在内。此外，本文通过使用以下方式引入了两种评估模式：i) 在测试集中使用所有测试的 X 光片或 ii) 仅使用测试集中的肺结核的 X 光片。有了这些指标，本文可以更全面地分析 CTD 系统的性能。

5. 实验和分析

5.1. 图像分类

本文将图像分类的评价结果总结在 Tab. 3 中。启用 ImageNet 预训练 [10] 时，Faster R-CNN [37] 的性能最好。在没有 ImageNet 预训练 [10] 时，SSD [27] 的表现最好。本文还观察到，ImageNet 预训练可以显著提高性能，但是 SSD 在不进行预训练的情况下获得了更好的性能。这可能是因为 SSD 采用了较浅主干的 VGGNet-16 [38]，它比 ResNet-50 [15] 更容易训练，其他方法使用 FPN [24] 。注意，如果 ImageNet 没有进行预训练，FCOS [40] 的训练会崩溃，因此本文在 Tab. 3 中不包含此设置的结果。

Faster R-CNN [37] 达到了 91.2% 的高灵敏度，这表明深度学习可以识别大多数肺结核 X 光片。特异性为 89.9%，这意味着 100 张非肺结核 X 光片中有 10.1 张被归类为肺结核。此外，在准确度方面，如 Sec. 3.3 所示，所有方法 (即不进行预训练的 SSD 方法和基于预训练 ResNet-50 的方法) 都比放射科医师的表现更好，因此基于深度学习的 CTD 是一个很有前途的研究领域。这一方向的未来进展有可能促进实用的 CTD 系统，从而帮助数百万肺结核患者。

5.2. 肺结核区域检测

在这里，本文报告了整个 TBX11K 测试集上及只使用测试集中肺结核 X 光片的各个方法的多种评价指标上的结果。由于非肺结核 X 光片中无肺结核病灶区域，如果评价时仅使用可以提供精确检测分析的肺结核 X 光片，同时使用所有 X 光片进行评估，这其中存在假阳性的非肺结核 X 光片的影响。在使用所有 X 光片进行评估时，本文使用了图像分类将分为非肺结核的 X 光片中的所有预测框进行过滤，而仅使用肺结核的 X 光片的过滤方式对于评价是无用的。由于 FCOS [40] 图像分类器的训练中若不预训练 ImageNet 就会崩溃，在这种情况下本文没有报告该方法下使用所有 X 光片的评估结果。

除 SSD [27] 外，ImageNet 预训练 [10] 可以提高检测性能。无论是否经过预处理，SSD 性能基本不变。SSD 在大多数情况下都获得了最佳的性能，除了使用所有的 X 光片和 ImageNet 预训练的结果，这种情况下 Faster R-CNN [37] 获得了最佳性能。虽然所有的方法都不能准确地发现陈旧性肺结核区域，但不考虑类别的肺结核的评价结果优于活动性肺结核，这意味着许多陈旧性肺结核目标被正确定位，但却被错误地归类为活动性肺结核。本文推测这是由

Method	是否预训练	主干网络	精确率 (Acc)	AUC (肺结核)	敏感率	特异率	平均准确率	平均召回率
SSD [27]	是	VGGNet-16	84.7	93.0	78.1	89.4	82.1	83.8
RetinaNet [25]		ResNet-50 w/ FPN	87.4	91.8	81.6	89.8	84.8	86.8
Faster R-CNN [37]		ResNet-50 w/ FPN	**89.7**	**93.6**	**91.2**	**89.9**	**87.7**	**90.5**
FCOS [40]		ResNet-50 w/ FPN	88.9	92.4	87.3	**89.9**	86.6	89.2
SSD [27]	否	VGGNet-16	**88.2**	**93.8**	**88.4**	89.5	**86.0**	**88.6**
RetinaNet [25]		ResNet-50 w/ FPN	79.0	87.4	60.0	**90.7**	75.9	75.8
Faster R-CNN [37]		ResNet-50 w/ FPN	81.3	89.7	72.5	87.3	78.5	79.9

表 3. TBX11K 测试集上的 X 光片图像分类结果。“是否预训练”表示是否对 ImageNet [10] 上的主干网络进行预训练。“主干网络”指每个基准方法使用的主干网络，其中 FPN 表示用于目标检测的特征金字塔网络 [24]。

方法名称	数据	是否预训练	主干网络	肺结核		活动性肺结核		陈旧性肺结核	
				AP^{bb}_{50}	AP^{bb}	AP^{bb}_{50}	AP^{bb}	AP^{bb}_{50}	AP^{bb}
SSD [27]	ALL	是	VGGNet-16	52.3	22.6	50.5	**22.8**	8.1	**3.2**
RetinaNet [25]			ResNet-50 w/ FPN	52.1	22.2	45.4	19.6	6.2	2.4
Faster R-CNN [37]			ResNet-50 w/ FPN	**57.3**	**22.7**	**53.3**	21.9	**9.6**	2.9
FCOS [40]			ResNet-50 w/ FPN	46.6	18.9	40.3	16.8	6.2	2.1
SSD [27]		否	VGGNet-16	**61.5**	**26.1**	**60.0**	**26.2**	**8.2**	**2.9**
RetinaNet [25]			ResNet-50 w/ FPN	20.7	7.2	19.1	6.4	1.6	0.6
Faster R-CNN [37]			ResNet-50 w/ FPN	21.9	7.4	21.2	7.1	2.7	0.8
SSD [27]	TB	是	VGGNet-16	68.3	**28.7**	**63.7**	**28.0**	**10.7**	4.0
RetinaNet [25]			ResNet-50 w/ FPN	**69.4**	28.3	61.5	25.3	10.2	**4.1**
Faster R-CNN [37]			ResNet-50 w/ FPN	63.4	24.6	58.7	23.7	9.6	2.8
FCOS [40]			ResNet-50 w/ FPN	56.3	22.5	47.9	19.8	7.4	2.4
SSD [27]		否	VGGNet-16	**69.6**	**29.1**	**67.0**	**29.0**	**9.9**	**3.5**
RetinaNet [25]			ResNet-50 w/ FPN	40.5	13.8	37.8	12.7	3.2	1.1
Faster R-CNN [37]			ResNet-50 w/ FPN	37.4	11.8	35.3	11.3	3.9	1.1
FCOS [40]			ResNet-50 w/ FPN	42.1	14.4	38.5	13.6	4.3	1.1

表 4. 在 TBX11K 测试集上的肺结核病灶区域检测结果。“数据”表示是使用所有测试 X 光片进行评估 (ALL)，还是只使用测试集的肺结核 X 光片进行评估 (TB) 。“肺结核”表示不考虑类别的肺结核检测结果。

于 TBX11K 中陈旧性肺结核 X 光片的数量有限造成的，其中只有 212 张陈旧性肺结核 X 光片，但有 924 张活动性肺结核 X 光片。

因此未来的研究应更加关注数据分布不均匀的问题。本文还发现，AP^{bb}_{50} 的性能通常比 AP^{bb} 的性能好得多。这意味着，虽然检测可以找到目标区域，但定位通常不是很准确。本文认为，定位肺结核边界框区域与定位自然目标区域有很大的不同。即使是有经验的放射科医师也不能轻易地确定肺结核的精确位置。因此，AP^{bb}_{50} 比 AP^{bb} 更重要，因为与肺结核目标 IoU 为 0.5 的预测框足以帮助放射科医生找到肺结核区域。

在 Fig. 2中，本文绘制各个方法的 PR 曲线用于检测误差分析。对于不考虑类别的肺结核检测评价，所有的方法都采用 ImageNet 预训练的策略。本文可以清楚地看到，IoU 的阈值从 0.75 到 0.5，所有的方法都有很大的改善。这表明，这些方法在较高的 IoU 阈值时，由于其较差的目标定位而评价分数较低。使用所有 X 光片时的“FN”区域远远大于仅使用肺结核 X 光片时的“FN”区域，这表明图像分类过滤了许多正确检测到的肺结核病灶，但本文认为这种过滤有助于提高整体检测性能。当使用所有的 X 光片进行

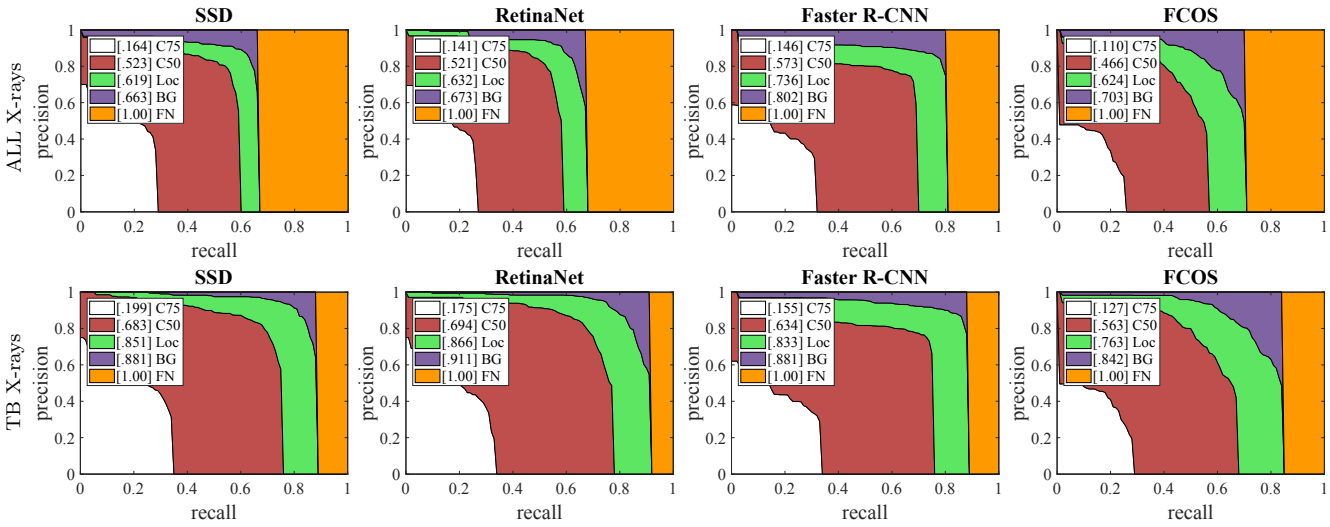

图 2. 在使用 ImageNet [10] 进行预训练的情况下，对不考虑类别的肺结核病灶区域检测的误差分析。第一行使用所有 X 光片进行评估，而第二行仅使用肺结核 X 光片。C50/C75: IoU 阈值为 0.5/0.75 时的 PR 曲线。Loc: IoU 为 0.1 下的 PR 曲线。BG: 去除背景误报 (FP)。FN: 删除其他未检测到的目标引起的错误。

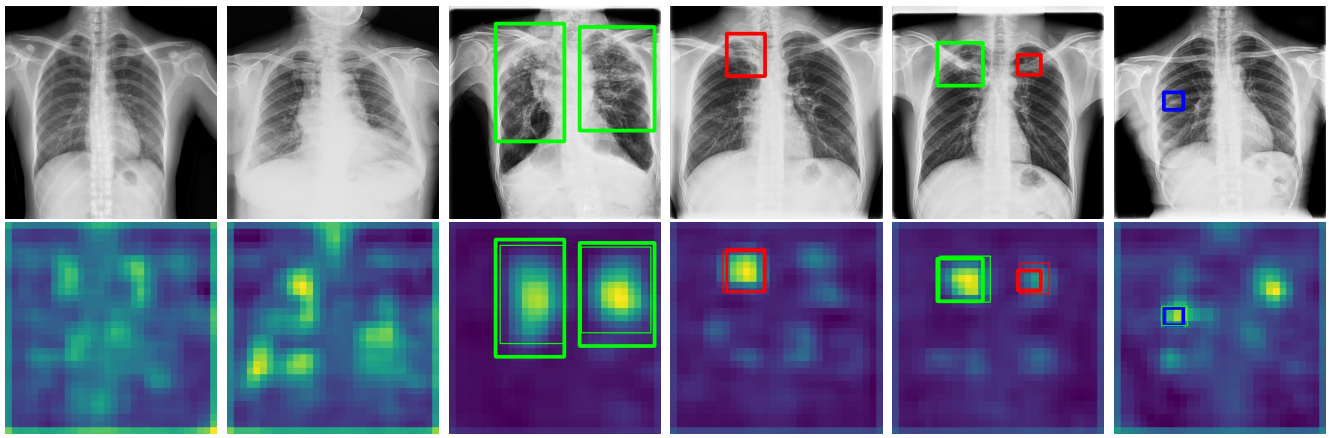

图 3. 从 X 光片中学习到的深层特征的可视化结果。所有的 X 光片都是从 TBX11K 测试集中随机选择的。对于表 2 中列出的每个类，本文给出一个例子。绿色、红色和蓝色盒子分别覆盖活动性肺结核、陈旧性肺结核和不确定型肺结核区域。粗线盒和细线盒分别表示真实边界框和 CTD 方法预测的边界框。

评估时，Faster R-CNN [37] 取得了最高的性能。当仅使用肺结核 X 光片进行评估时，RetinaNet [25] 似乎能取得更好的性能。结合图像的分类和肺结核区域的检测，本文可以推断出这些基线方法在不同的方面表现出了各自的优势。

5.3. 可视化

为了解卷积神经网络从不同的 X 光片中学习了到什么，本文以 RetinaNet [25] 为主干的 1/32 比例可视化特征图。具体来说，本文使用主成分分析将特征图的通道缩减为一个通道。此单通道图被转换为热图以便进行可视化。本文将结果显示在 Fig. 3中。健康病例的可视化结果是不规则的，而非肺结核患者的可视化结果有一些亮点，可能是病变。对于肺结核病例，可视化图中的高亮显示与标注的 TB 区域一致。

6. 结论

肺结核是一种主要的传染病，早期诊断对肺结核的治疗和预防很重要。不幸的是，肺结核诊断仍然是一个重大挑战。使用金标准对肺结核进行最终检验需要几个月的时间，而且在许多发展中国家和资源紧张的社区是不可能的。受深度学习的成功启发，基于深度学习的 CTD 是一个很有前途的研究方向。然而，数据的缺乏阻碍了深度学习为 CTD 带来进步。本文构建了一个带有边界框标注的大规模肺结核数据集 TBX11K，能够训练用于肺结核诊断的深度 CNNs。TBX11K 也是肺结核区域检测的第一个数据集。通过进一步提出一些基准和评估指标，本文为 CTD 建立了一个初始基准。这一新的 TBX11K 数据集和基准有望推动 CTD 的研究，并有望与新的强大的深度网络 [12] 一起设计更好的 CTD 系统。

致谢 本研究项目由 No.2018AAA0100400 批准的新一代人工智能重大项目、国家自然科学基金 (61922046)、国家青年人才支持计划、天津市自然科学基金 (17JCJQJC43700, 18ZXZNGX00110) 支持。

参考文献

[1] P Andersen, ME Munk, JM Pollock, and TM Doherty. Specific immune-based diagnosis of tuberculosis. *The Lancet*, 356(9235):1099–1104, 2000. 1, 4, 5

[2] Aliya Bekmurzayeva, Marzhan Sypabekova, and Damira Kanayeva. Tuberculosis diagnosis using immunodominant, secreted antigens of mycobacterium tuberculosis. *Tuberculosis*, 93(4):381–388, 2013. 1, 4, 5

[3] Anna Bosch, Andrew Zisserman, and Xavier Munoz. Representing shape with a spatial pyramid kernel. In *ACM International Conference on Image and Video Retrieval*, pages 401–408. ACM, 2007. 3

[4] Yuri Boykov and Gareth Funka-Lea. Graph cuts and efficient N-D image segmentation. *International Journal of Computer Vision*, 70(2):109–131, 2006. 3

[5] Sema Candemir, Stefan Jaeger, Kannappan Palaniappan, Jonathan P Musco, Rahul K Singh, Zhiyun Xue, Alexandros Karargyris, Sameer Antani, George Thoma, and Clement J McDonald. Lung segmentation in chest radiographs using anatomical atlases with nonrigid registration. *IEEE Transactions on Medical Imaging*, 33(2):577–590, 2013. 1, 2, 3

[6] Arun Chauhan, Devesh Chauhan, and Chittaranjan Rout. Role of gist and phog features in computer-aided diagnosis of tuberculosis without segmentation. *PloS One*, 9(11):e112980, 2014. 1, 2, 3, 6

[7] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, et al. Mmdetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019. 6

[8] Ming-Ming Cheng, Yun Liu, Qibin Hou, Jiawang Bian, Philip Torr, Shi-Min Hu, and Zhuowen Tu. HFS: Hierarchical feature selection for efficient image segmentation. In *European Conference on Computer Vision*, pages 867–882, 2016. 3

[9] Ming-Ming Cheng, Yun Liu, Wen-Yan Lin, Ziming Zhang, Paul L Rosin, and Philip HS Torr. BING: Binarized normed gradients for objectness estimation at 300fps. *Computational Visual Media*, 5(1):3–20, 2019. 6

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 3, 6, 7, 8, 9

[11] Neel R Gandhi, Paul Nunn, Keertan Dheda, H Simon Schaaf, Matteo Zignol, Dick Van Soolingen, Paul Jensen, and Jaime Bayona. Multidrug-resistant and extensively drug-resistant tuberculosis: a threat to global control of tuberculosis. *The Lancet*, 375(9728):1830–1843, 2010. 1

[12] Shang-Hua Gao, Ming-Ming Cheng, Kai Zhao, Xin-Yu Zhang, Ming-Hsuan Yang, and Philip Torr. Res2Net: A new multi-scale backbone architecture. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. 2, 10

[13] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *IEEE International Conference on Computer Vision*, pages 2961–2969, 2017. 2

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-

level performance on imagenet classification. In *IEEE International Conference on Computer Vision*, pages 1026–1034, 2015. 2

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 2, 6, 7

[16] Xiaowei Hu, Yun Liu, Kai Wang, and Bo Ren. Learning hybrid convolutional features for edge detection. *Neurocomputing*, 313:377–385, 2018. 2

[17] Sangheum Hwang, Hyo-Eun Kim, Jihoon Jeong, and Hee-Jin Kim. A novel approach for tuberculosis screening based on deep convolutional neural networks. In *Medical Imaging 2016: Computer-Aided Diagnosis*, volume 9785, page 97852W. International Society for Optics and Photonics, 2016. 1, 3

[18] Stefan Jaeger, Sema Candemir, Sameer Antani, Yì-Xiáng J Wáng, Pu-Xuan Lu, and George Thoma. Two public chest X-ray datasets for computer-aided screening of pulmonary diseases. *Quantitative Imaging in Medicine and Surgery*, 4(6):475–477, 2014. 2, 3, 6

[19] Stefan Jaeger, Alexandros Karargyris, Sema Candemir, Les Folio, Jenifer Siegelman, Fiona Callaghan, Zhiyun Xue, Kannappan Palaniappan, Rahul K Singh, Sameer Antani, et al. Automatic tuberculosis screening using chest radiographs. *IEEE Transactions on Medical Imaging*, 33(2):233–245, 2013. 1, 2, 3

[20] Alexandros Karargyris, Jenifer Siegelman, Dimitris Tzortzis, Stefan Jaeger, Sema Candemir, Zhiyun Xue, KC Santosh, Szilárd Vajda, Sameer Antani, Les Folio, et al. Combination of texture and shape features to detect pulmonary abnormalities in digital chest X-rays. *International Journal of Computer Assisted Radiology and Surgery*, 11(1):99–106, 2016. 2, 3

[21] Anastasios Konstantinos. Testing for tuberculosis. *Australian Prescriber*, 33(1):12–18, 2010. 1

[22] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 1097–1105, 2012. 3

[23] Paras Lakhani and Baskaran Sundaram. Deep learning at chest radiography: automated classification of pulmonary tuberculosis by using convolutional neural networks. *Radiology*, 284(2):574–582, 2017. 3

[24] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2117–2125, 2017. 7, 8

[25] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *IEEE International Conference on Computer Vision*, pages 2980–2988, 2017. 2, 6, 8, 9

[26] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision*, pages 740–755, 2014. 7

[27] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. SSD: Single shot multibox detector. In *European Conference on Computer Vision*, pages 21–37, 2016. 2, 6, 7, 8

[28] Yun Liu, Ming-Ming Cheng, Deng-Ping Fan, Le Zhang, JiaWang Bian, and Dacheng Tao. Semantic edge detection with diverse deep supervision. *arXiv preprint arXiv:1804.02864*, 2018. 2

[29] Yun Liu, Ming-Ming Cheng, Xiaowei Hu, Jia-Wang Bian, Le Zhang, Xiang Bai, and Jinhui Tang. Richer convolutional features for edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(8):1939–1946, 2019. 2

[30] Yun Liu, Peng-Tao Jiang, Vahan Petrosyan, Shi-Jie Li, Jiawang Bian, Le Zhang, and Ming-Ming Cheng. DEL: Deep embedding learning for efficient image segmentation. In *International Joint Conference on Artificial Intelligence*, pages 864–870, 2018. 3

[31] Yun Liu, Yu-Huan Wu, Yunfeng Ban, Huifang Wang, and Ming-Ming Cheng. Rethinking computer-aided tuberculosis diagnosis. In *IEEE CVPR*, 2020. 1

[32] UK Lopes and João Francisco Valiati. Pre-trained convolutional neural networks as feature extractors for tuberculosis detection. *Computers in Biology and Medicine*, 89:135–143, 2017. 1, 2, 3

[33] Aude Oliva and Antonio Torralba. Building the gist of a scene: The role of global image features in recognition. *Progress in Brain Research*, 155:23–36, 2006. 3

[34] World Health Organization. Global tuberculosis report 2015. http://apps.who.int/iris/bitstream/10665/191102/1/9789241565059_eng.pdf, 2015. 1

[35] World Health Organization. Global tuberculosis report 2017. https://www.who.int/tb/publications/global_report/gtbr2017_main_text.pdf, 2017. 1

[36] World Health Organization et al. Chest radiography in tuberculosis detection: summary of current who recommendations and guidance on programmatic approaches. Technical report, World Health Organization, 2016. 1

[37] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*, pages 91–99, 2015. 2, 6, 7, 8, 9

[38] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015. 2, 6, 7

[39] Yi Sun, Yuheng Chen, Xiaogang Wang, and Xiaoou Tang. Deep learning face representation by joint identification-verification. In *Advances in Neural Information Processing Systems*, pages 1988–1996, 2014. 2

[40] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. FCOS: Fully convolutional one-stage object detection. In *IEEE International Conference on Computer Vision*, pages 9627–9636, 2019. 2, 6, 7, 8

[41] MRA Van Cleeff, LE Kivihya-Ndugga, H Meme, JA Odhiambo, and PR Klatser. The role and performance of chest x-ray for the diagnosis of tuberculosis: A cost-effectiveness analysis in nairobi, kenya. *BMC Infectious Diseases*, 5(1):111, 2005. 1

[42] Ziming Zhang, Yun Liu, Xi Chen, Yanjun Zhu, Ming-Ming Cheng, Venkatesh Saligrama, and Philip HS Torr. Sequential optimization for efficient high-quality object proposal generation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(5):1209–1223, 2018. 6