通过迭代细化和统计结果验证，高效训练人机协作的深度学习模型

Manuel Zahn[a]，**Douglas P. Perrin，哲学博士**[a,b,*]哈佛大学工程与应用科学学院，美国马萨诸塞州剑桥，邮编 02138[b]波士顿儿童医院心脏外科、哈佛医学院，美国马萨诸塞州波士顿，邮编 02115

arXiv:2304.00990v1 [cs.CV] 3 Apr 2023

摘要。

图像标注是将深度学习应用于医疗数据时面临的最大挑战之一。目前的流程既耗时又成本高昂，因此限制了这项技术的广泛应用。此外，验证测得的性能提升是否显著，对于选出最佳模型至关重要。本文展示了一种创建分割标注的方法，这是超声成像机器学习流程中数据清理的必要环节。我们提出一种四步法，利用自动生成的训练数据和快速的人工视觉检查，在降低时间、精力和成本的同时提高模型准确率。我们还展示了通过多次重复实验来进行统计分析的方法。质量欠佳的自动生成真实标注数据与快速视觉检查，能够高效地训练初始基准模型；随后再使用少量成本较高的人工生成真实标注数据对其进行细化。我们在心脏超声分割任务中演示了该方法，并去除了背景数据，包括静态受保护健康信息。通过多次重复实验，并对性能分布进行学生t检验，验证了结果的显著性。简单阈值算法的初始分割准确率为 92%，经改进后达到 98%。先使用性能较差的算法进行预训练，再辅以少量高质量数据，便可达到或超越使用复杂算法训练的模型性能。引入针对深度学习模型的统计显著性分析，有助于验证所测得的性能提升。该方法以经济高效、快速的方式构建高准确率模型，同时尽量降低获取高质量训练数据所需的成本和精力。

关键词： 标注、深度学习、迁移学习、超声、分割、验证。

*Douglas P. Perrin，douglas.perrin@childrens.harvard.edu

1 引言

近年来，深度学习技术广受欢迎。尽管取得了显著成果，但其开发过程仍然成本高昂且耗时。因此，研究重点在于加快模型训练、缩小深度学习模型的规模，并减少所需的训练数据量。Han 等人在 2016 年提出了一种早期方法。他们利用网络剪枝，在准确率不受影响的情况下，将 AlexNet[1] 所需的存储空间减少了 35 倍，并将 VGG-16[2] 的规模缩小了 49 倍。[3] Frankle 和 Carbin 提出了“彩票假设”，旨在找出比原始网络学习速度更快、测试时达到更高

的准确率。[4] Choi 等人提出了“数据回响”这一概念。在通过网络读取训练数据时，该方法将 ResNet-50[5] 在 ImageNet 上[6] 的实际运行时间缩短了 3.25 倍。[7]

除了模型训练外，数据标注也是开发监督式深度学习模型时最耗时、成本最高的任务之一。对于许多医疗应用，模型还需要达到很高的准确率。专家通常要花数小时进行图像标注。此外，在训练出初始基线模型之前，往往需要标注大量图像。目前，对于医疗机器学习问题而言，尚未明确医学影像数据集需要达到怎样的数据量和图像质量才算充足。[8] 这些标注方面的挑战限制了深度学习技术的广泛应用，也使模型开发过程变得漫长且成本高昂。

本文重点探讨如何建立一种流程，以减少创建分割掩膜所需的昂贵训练数据量。我们以单锥超声图像为例展示了这一流程，并可将其用作去除烧录在图像中的患者健康数据的工具。我们提出了一种四步法，可大幅减少创建精准深度学习分割模型所需的高质量人工标注数量。该方法利用质量较差的自动生成标签和快速人工视觉筛选来实现这一目标。此流程缩短了模型开发时间，并使那些难以且昂贵地获取高质量标签的应用场景成为可能。我们还展示了如何改进简单算法，使其性能达到复杂算法的水平，从而减少在复杂初始分割算法上投入大量开发时间的必要性。此前，人们曾通过使用预训练模型权重来减少对大量昂贵训练数据的需求。许多应用会利用在 ImageNet[9] 等大型数据集上训练的权重，并将其应用于图像领域相似的其他问题。当处理医疗数据时，这种方法并不合理，因为大多数大规模图像数据集都不包含医学图像数据。

例如，许多自然图像中常见的清晰边缘、纹理和色彩空间渐变等特征，在许多类型的医学图像中都并不存在。对于大规模自然图像数据集的预训练权重在医学图像任务中的应用，这种显著领域差异所带来的影响尚未得到充分表征。

因此，在我们的工作中，我们将从头开始训练基础模型，然后使用一个规模较小、由人工分割的真实标注数据集进行额外训练，以此对其进行优化。R. Barth 等人提出了一个用于农业分割的类似概念。在该研究中，研究人员使用合成甜椒图像数据集训练深度学习模型，然后通过在一个规模较小、由人工分割的真实世界图像数据集上进行额外训练，进一步优化模型性能。[10] 与 R. Barth 等人的方法不同，我们将利用不同复杂程度的计算机视觉算法，并快速进行人工筛选，以创建用于基础模型的训练集，从而取代耗时的合成图像制作过程。总的来说，这种方法能够以快捷、低成本（包括人力和计算成本）的方式创建深度学习训练数据。

本文所述方法可作为其他机器学习方法的前置步骤。对医学图像数据进行掩膜处理，也有助于去除定位标记及其他背景物体。这些因素可能造成预测偏差；如果部署在临床环境中，可能会给深度学习模型带来灾难性后果。因此，数据准备和预处理是确保训练数据不含偏差的关键任务。通常无法获取不含能表明图像来源特征的原始图像数据。例如，图像上叠加的制造商标识、产品名称、饱和度条的特定位置、视觉跟踪标记，以及患者姓名或年龄等受保护健康信息数据。在使用包含多个数据来源的数据集训练深度学习模型时，数据中的这类视觉标记可能会

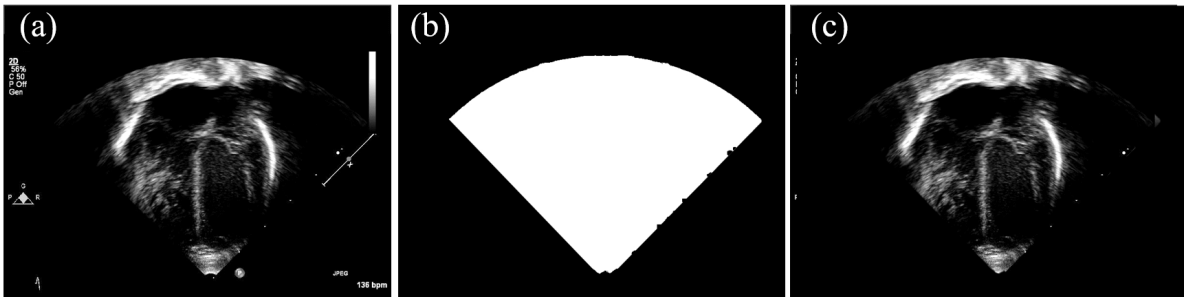

图 1 (a)：包含非回声杂波的输入图像，其中可能含有受保护健康信息（PHI）；(b)：仅包含超声图像数据的二值掩膜；(c)：应用掩膜后不含任何潜在偏差和受保护健康信息的输入图像。

导致严重的模型偏差，因此必须加以处理。

本文将以心脏超声背景去除任务展示我们的方法。所提出的概念也可应用于胎儿及其他单锥形超声图像数据。在临床超声图像中，患者信息、饱和度条、心电图（EKG）等静态和动态元素会与超声数据合并在同一帧中（图 1，a）。由于这种偏差，锥形区域分割是超声数据深度学习模型的一项必要预处理步骤。为此，我们创建并应用二值掩膜（图 1，b）到输入图像（图 1，c）。我们将使用复杂度各异的计算机视觉算法生成基础模型的训练集，以比较它们采用所述方法后性能的提升情况。昂贵的手动分割数据仅用于模型精炼步骤。

2 方法

2.1 数据收集

本文的数据集来自波士顿儿童医院和儿科心脏网络。它包含从美国各地儿童医院收集的 1,666 个心脏序列。每个序列由 10 - 120 帧组成，至少显示一次心跳。这些帧以

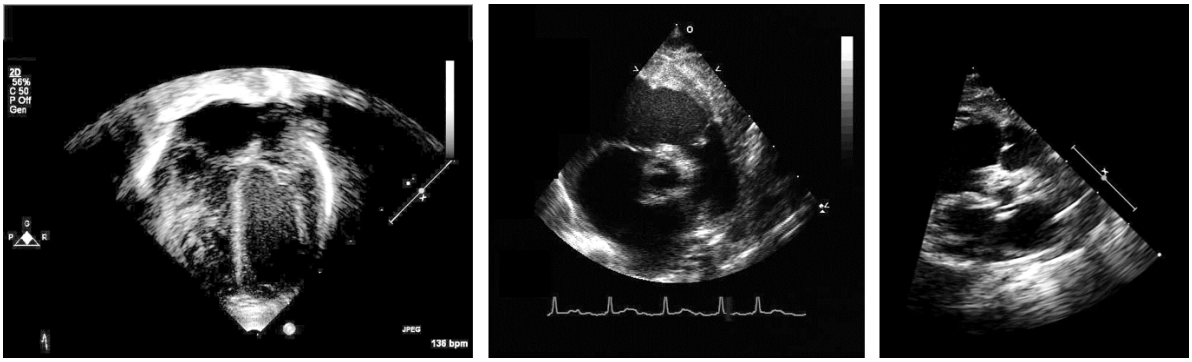

图 2 不同心跳序列中的一些示例帧，展示了数据集中旋转、尺度和结构的变化。

便携式网络图形（PNG）格式提供，且未进行分割或标注。在这些序列中，超声锥体的尺度、位置、旋转和饱和度各不相同。有些锥体被遮挡或截断，而在某些序列中，背景物体（静态和动态）覆盖了锥体结构。图像序列有 7 种不同尺寸，最大为 800 x 600 像素，最小为 480 x 430 像素。部分精选图像见图 2。

2.2 模型精炼方法

我们的模型优化方法由四个步骤组成。第一步，我们使用计算机视觉对超声锥体进行分割，从而创建初始训练掩膜。我们假设，这些掩膜算法的输出可用于训练表现平平的基础模型，再利用一小组成本高昂的人工分割数据，将其优化至更高的准确率。在此过程中，较简单和较复杂的初始计算机视觉分割算法之间的方差会降低。这种优化方法的优势在于，能够减少计算机视觉算法开发、真值生成或两者所需的工作量。在本研究中，我们将使用简单阈值分割、填充式阈值分割和凸包算法的结果，分别训练三个基础模型。从数据集中随机选取 33 张图像作为测试集。这些图像经过人工

分割，以创建一个能够代表整个数据集的真实标注测试集（标记为：代表性测试集）。该测试集将用于评估模型准确率，其中的一个子集将用于执行精炼步骤。第二步是快速手动筛选标签掩膜的质量。手动筛选利用人类视觉系统的效率来区分良好和较差的分割结果。手动进行简单的“纳入／排除”筛选，速度远快于手工分割。视觉检查中被判定为良好的图像随后将用作基础模型的训练集。从手动筛选中被判定为较差的图像里，再随机选取另外 33 张不同的图像作为测试集。这些图像经过手工分割，以创建一个仅包含较差图像的测试集（标记为：Bad Mask 测试集）。该测试集可用于评估计算机视觉算法自动分割失败时模型在相应案例中的表现。第三步，使用被判定为良好的掩膜训练基础模型。模型在各测试集上的表现，可用于比较所提出的方法带来的模型改进。最后，在第四步中，使用“代表性测试集”中一部分手工分割的图像来优化基础模型。最终的模型性能评估通过在“Bad Mask 测试集”上进行评估得出。图 3展示了完整的模型优化方法。我们进行了六次建模实验，并提出利用对所得模型性能分布进行统计分析，以衡量实验结果的显著性。

2.3 结果

2.3.1 使用计算机视觉算法创建数据集

我们使用不同的计算机视觉算法创建了三个未经排序的初始训练数据集。为了移除静态背景物体，我们利用给定图像的序列特性。通过对相邻帧进行差分，仅保留发生变化的图像部分。一个序列中的 N 帧

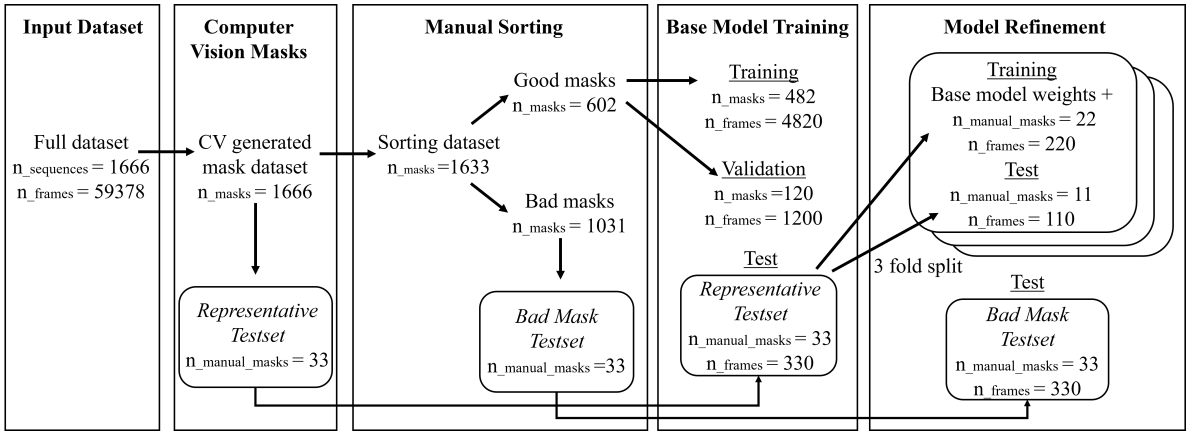

图 3 完整的模型优化方法。

（图4，a）因此会生成 N-1 张差分图像。随后，可通过计算这 N-1 张差分图像的均值，将它们合成为一张图像（图4，b）。接着，以这张平均差分图像为基准，依次创建本文性能比较中使用的三种掩膜类型。最简单的分割掩膜通过对基准图像应用自适应阈值获得（图4，c）。由于该掩膜是根据合成后的均值图像创建的，因此适用于整个图像序列。下一步，对阈值掩膜应用泛洪填充算法。由此得到的掩膜会填补结构中的孔洞，并生成更精确的“填充阈值”掩膜（图4，d）。最后，应用凸包算法，进一步提升结果（图4，e）。这些步骤在整个数据集上执行，从而生成一组用于基础模型训练的分割掩膜。由于无需进行手工分割或创建合成超声图像，因此成本相对较低。对“代表性测试集”应用相同的计算机视觉算法后，可计算出阈值掩膜、填充阈值分割掩膜和凸包掩膜的初始准确率分别为 91.87%、95.74% 和 97.24%。准确率定义为图像中所有真阳性（TP）和真阴性（TN）像素之和除以图像中的像素总数。

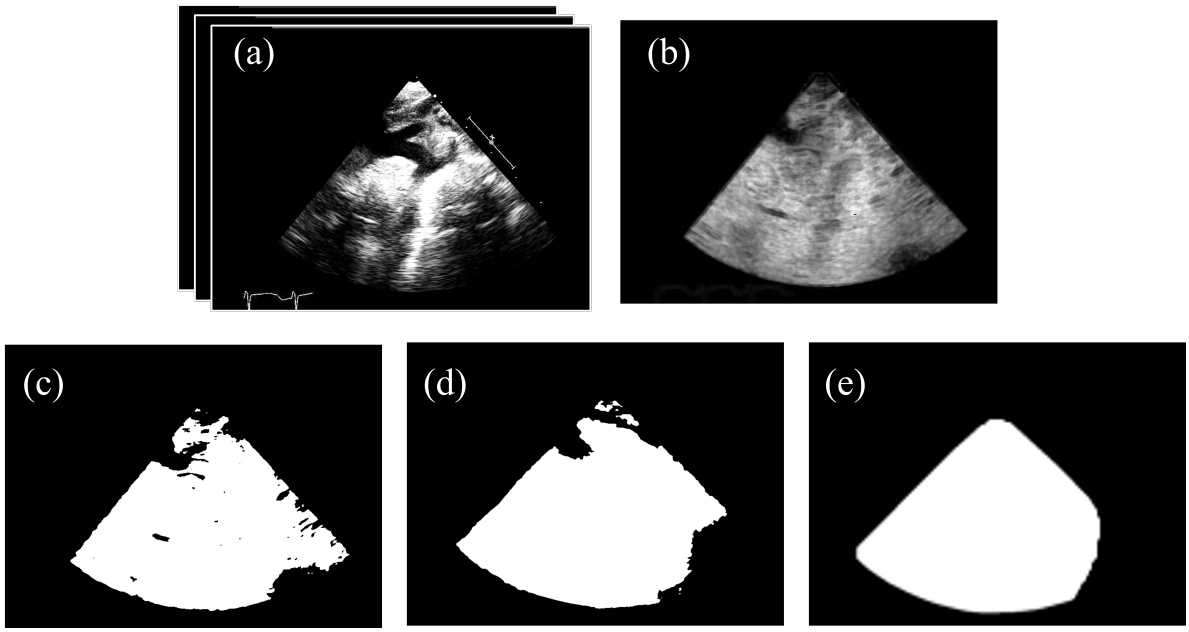

图 4 训练掩膜的生成流程：(a) n 个原始心脏帧；(b) n-1 幅差分图像的均值；(c) 通过自适应阈值化生成二值掩膜；(d) 应用泛洪填充算法后的掩膜；(e) 应用凸包算法后的掩膜

2.3.2 图像的手动排序

排序通过将所有掩膜分为两类，利用了人类的视觉评估。覆盖超声锥形区域大部分的掩膜被视为良好，并纳入训练集。包含部分背景，或掩膜区域遗漏了超声锥形区域大部分内容的掩膜，则被视为自动分割失败并予以剔除。事实证明，人工评估比创建手工分割的真实标注掩膜或合成超声数据快得多，同时仍能发挥人类的判断能力。手动分割单张图像平均需要三分钟，因此耗时更多。对于全部 1666 张图像，需要 4998 分钟，即 83.3 小时的人工分割时间。相比之下，对计算机视觉算法在整个数据集上生成的掩膜进行纳入或剔除分类，不到 20 分钟便完成了。分类后，序列中的 602 个掩膜被判定为良好，并保留为基础模型的训练集，仅占初始数据集的 37%。

因此，使用传统计算机视觉算法生成的分割标注中，有 63% 未通过视觉评估。

2.3.3 基础模型的训练

基础模型使用不含不良掩膜的数据集进行训练。我们选择 U-Net 结构[11] 作为模型架构。它是一种专为生物医学图像分割设计的卷积神经网络，由收缩路径和扩张路径组成，整体呈 U 形结构。我们的实验在哈佛大学 FAS 研究计算集群上运行，使用了 4 个图形处理器，所有核心共用一个 32GB 内存池。训练采用均方误差 Adam 优化器[12]，学习率为 1e-5，批量大小为 8 张图像。我们留出 20% 的输入数据用于验证，图像大小为 256x256 像素。由于每个序列只有一个掩膜，因此可以使用同一序列中的多张图像，并配以相同的掩膜进行训练。我们为每个序列随机选取十帧，每帧都使用对应的掩膜进行模型训练。每隔 10 个训练轮次，便在人工分割的“代表性测试集”上评估学习曲线。实验共运行六次，每次运行都为每个序列选取不同的十个随机帧。这样便可以通过比较各次运行的模型性能分布来进行统计显著性计算。曲线显示了三种输入掩膜类型在六次运行中的平均准确率和方差。图 5展示了在测试集上评估的基础模型训练准确率及方差。六次模型运行的方差在初期较高，但随着训练轮次增加而逐渐减小。经过 100 个训练轮次后，阈值分割、填充式阈值分割和凸包掩膜的平均准确率分别达到 94.99%、97.71% 和 98.49%。与基于计算机视觉的算法相比，准确率分别提高了 3.12%、1.97% 和 1.25%。这也表明，不同输入掩膜类型之间仍存在较大的性能差异

，最简单的计算机视觉阈值分割输入掩膜与更复杂的凸包输入掩膜之间相差 3.5%。我们推测，性能仍有进一步提升的空间，同时也可以降低这种由输入掩膜类型导致的性能方差。

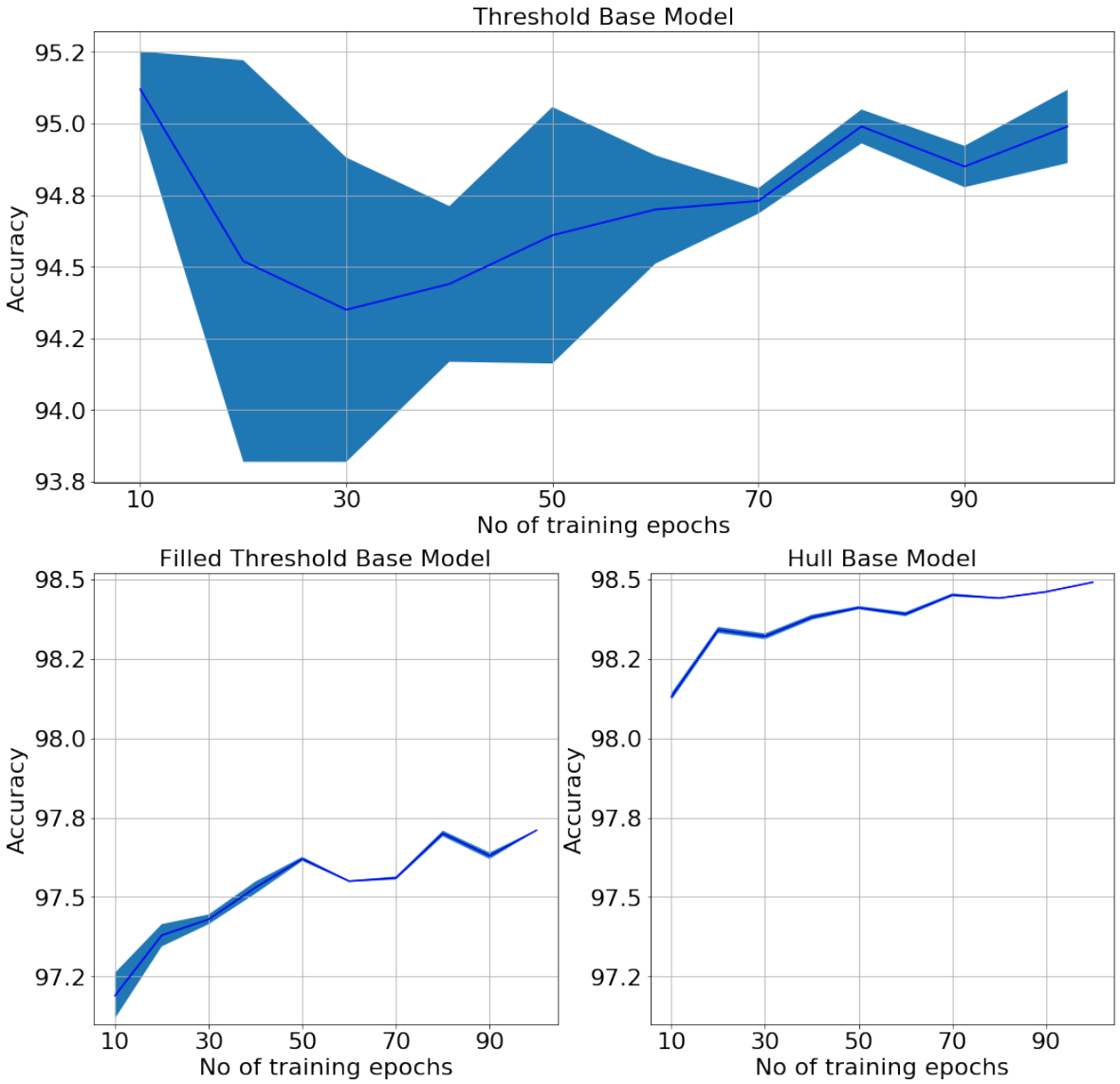

图 5 在“代表性测试集”上评估的基础模型每十个训练轮次的平均训练准确率和方差。上排：阈值模型；下排左侧：填充阈值模型；下排右侧：凸包模型

2.3.4 使用手工分割数据集进行模型精炼

基础模型的表现已经优于计算机视觉算法。我们假设，只需少量准确的手工分割数据集，就能进一步提升模型性能。应通过对多次训练运行的性能分布进行统计分析来验证这一点。该方法可使较简单的计算机视觉算法达到与更复杂的算法或需要大量昂贵训练数据的算法相当的准确率。其思路是使用测试集的一部分来精炼基础模型，同时用剩余测试集评估新模型。因此，“代表性测试集”被分为三部分：三分之二通过额外训练来精炼基础模型，三分之一用于评估性能。对所有三种可能的组合（第 1 折、第 2 折和第 3 折）重复这一过程。每项实验使用不同的随机帧进行六次训练，从而计算各训练轮次的均值和方差。结果见图6。由于每个序列中取十帧，因此每个训练集包含 220 个训练样本，其中 20% 用于验证，另有 110 个测试样本用于评估性能。这也使得能够按照学生 t 检验的描述，通过计算 p 值比较两个模型，[13] 使用每个模型的六次运行结果。表1展示了基础模型和三种细化模型的平均准确率及 95% 置信区间。评估两个模型性能差异的统计显著性时，采用 Holm-Bonferroni 校正来控制族错误率。如果排序后的 p 值小于对应的 $\alpha_{Holm-Bonferroni}$ 值，则所有相对于基础模型的改进均具有统计显著性。 $\alpha_{Holm-Bonferroni}$这些实验的比较值依次为 8.3e-03、1.00e-02、1.25e-02、1.67e-02、2.50e-02 和 5.00e-02；符合统计显著性标准的比较以星号标记

。结果见（表2）。除凸包掩膜的第三次细化外，所有情况下细化模型的平均准确率均高于基础模型。除填充阈值的第三次细化外，所有测得的变化在统计上均具有显著性，其 p 值低于对应的 $\alpha_{Holm-Bonferroni}$。总体而言，比较细化模型时，初始计算机视觉掩膜算法之间的性能差异大幅缩小。阈值算法与凸包基础模型之间原本 3.5% 的准确率差距，在所有折中均缩小至 1% 以下，且第二次细化时仅为 0.15%。这表明，细化方法能够以开发时间更短的简单计算机视觉算法，达到与更复杂方法相近的性能准确率。

表 1 模型准确率均值（n=6）和 95%置信区间

模型	阈值		填充阈值		Hull	
	平均准确率	95%置信度	平均准确率	95%置信度	平均准确率	95%置信度
Base	94.99	(94.58, 95.41)	97.71	(97.68, 97.75)	98.49	(98.44, 98.54)
第1个参考项	98.14	(98.09, 98.2)	98.47	(98.41, 98.54)	98.7	(98.62, 98.78)
第2个参考项	98.87	(98.84, 98.9)	98.93	(98.92, 98.95)	99.02	(98.97, 99.06)
第3个参考项	97.6	(97.49, 97.71)	97.78	(97.67, 97.9)	98.28	(98.2, 98.36)

表 2 不同模型的统计显著性评估。

模型比较	阈值	填充阈值	Hull
基准与第1个参考项	＜1e-08*	＜1e-08*	1.58e-04*
基准 vs. 第2个参考项	＜1e-08*	＜1e-08*	＜1e-08*
基准与第3个参考项	2.22e-08*	1.9e-01	2.12e-04*
第1个参考项与第2个参考项	＜1e-08*	＜1e-08*	4.0e-06*
第1个参考项 与 第3个参考项	4.81e-07*	1.00e-07*	2.42e-06*
第2个参考项 vs. 第3个参考项	＜1e-08*	＜1e-08*	＜1e-08*

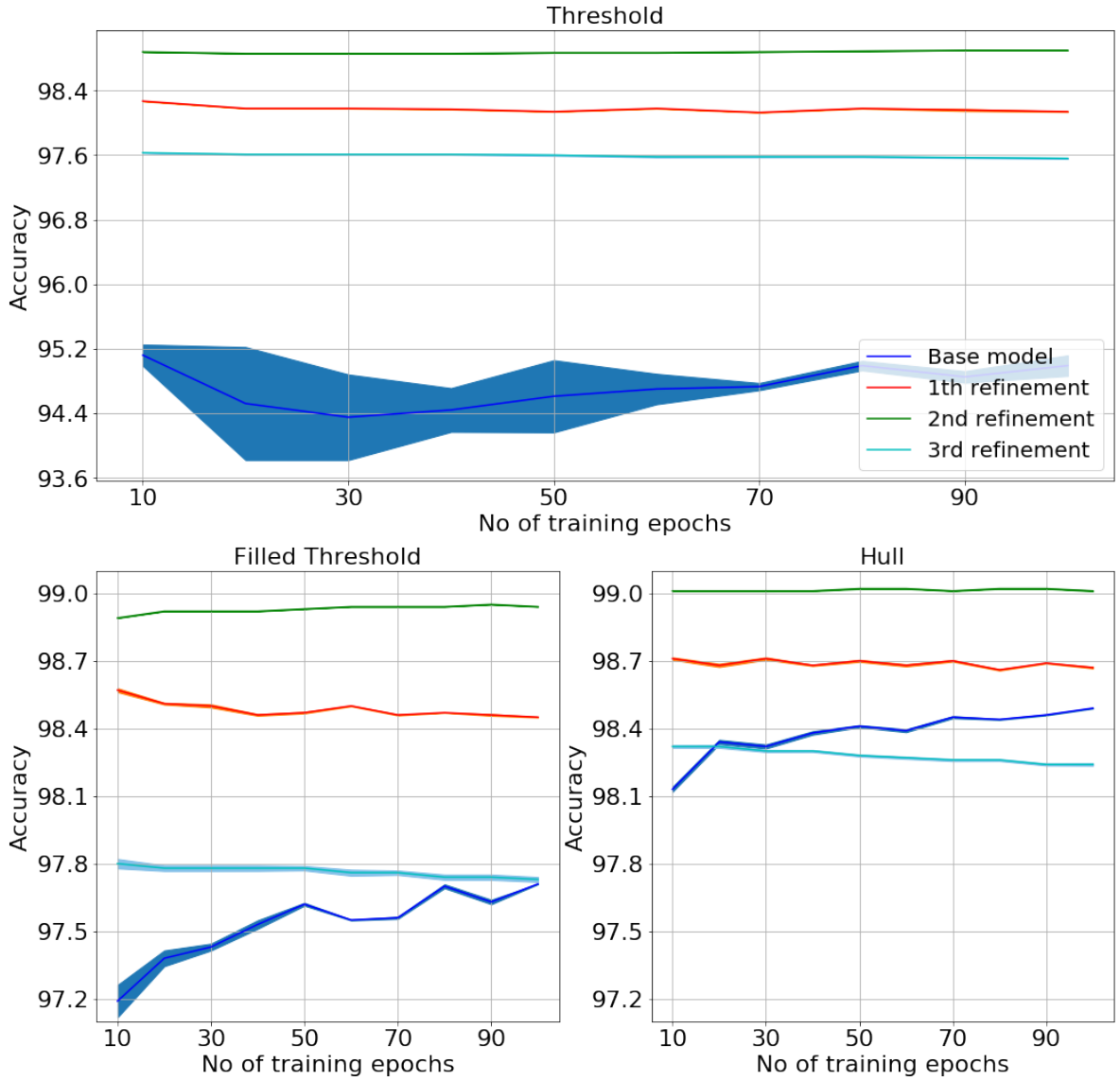

图 6 在“代表性测试集”上评估的基准模型和三个优化划分每十个训练轮次的平均训练准确率及方差。上排：阈值模型；下排左侧：填充阈值模型；下排右侧：凸包模型

2.3.5 坏图像测试集评估

我们评估了深度学习模型在自动分割未能生成可接受结果的图像上的表现。为此，我们使用了“Bad Mask 测试集”；该测试集包含 33 张图像，这些图像在细化方法的第二步中被判定为不合格（图 3）。同样，我们从每个序列中随机选取 10 帧，创建了一个包含 330 张图像的坏图像测试集。随后

将基准模型和细化模型与该测试集进行比较评估。

考虑到数据集仅包含最初分割效果不佳的图像，整体表现良好。基础模型的平均准确率分别为：阈值模型 93.16%、填充阈值模型 96.32%、凸包模型 97.93%（表 3）。除第三个凸包细化模型外，所有细化模型的准确率都更高。第二折细化模型的平均准确率最高，阈值模型为 97.56%、填充阈值模型为 97.78%、凸包模型为 98.13%。在评估统计显著性时，阈值模型和填充阈值模型相较于基础模型的所有改进均具有显著性。对于凸包模型，第二次细化带来的改进在 Holm-Bonferroni 校正后仍具有统计显著性（表 4）。同样可以看出，与细化模型相比，最初计算机视觉掩膜算法之间的差异大幅缩小。阈值模型与凸包基础模型之间最初 4.77% 的准确率差异，在所有折中均缩小至 1% 以下，第二次细化时更是降至仅 0.57%。

Tab表 3 “Bad Mask 测试集”上的模型准确率均值（n=6）及95%置信区间 ”

模型	阈值		填充阈值		Hull	
	平均准确率	95%置信度	平均准确率	95%置信度	平均准确率	95%置信度
Base	93.16	(92.69, 93.63)	96.32	(96.2, 96.45)	97.93	(97.8, 98.06)
第1个参考项	97.18	(97.11, 97.25)	97.51	(97.45, 97.58)	98.02	(97.94, 98.1)
第2个参考项	97.56	(97.48, 97.64)	97.78	(97.7, 97.86)	98.13	(98.06, 98.2)
第3个参考项	96.91	(96.75, 97.08)	97.23	(97.12, 97.35)	97.8	(97.69, 97.92)

2.3.6 图像去标识化的视觉评估

为评估模型从输入图像中移除文本、定位标记及其他背景数据的能力，我们还对模型生成的分割掩膜进行了目视评估。如果图像中仍有可见文本或可能属于非超声受保护健康信息的内容，则应计为图像去标识化失败

表 4 不同模型的统计显著性评估。

模型比较	阈值	填充阈值	Hull
基准 vs. 第1个参考项	＜1e-08*	＜1e-08*	1.53e-01
基准 vs. 第2个参考项	＜1e-08*	＜1e-08*	5.06e-03*
基准与第3个参考项	＜1e-08*	6.95e-08*	9.21e-02
第1个参考项与第2个参考项	4.42e-06*	5.94e-05*	4.0e-06*
第1个参考项 vs. 第3个参考项	2.9e-03*	2.70e-04*	1.95e-02
第2个参考项与第3个参考项	3.48e-06*	1.37e-06*	8.08e-05*

。如果图像仅包含超声数据，且分割后没有可见文本或可能属于受保护健康信息的内容，则无论分割本身的质量如何，都应计为图像去标识化成功。

在“Bad Mask 测试集”的视觉背景移除评估中，所有模型（基准模型、第1次、第2次和第3次细化）表现同样出色。成功去除了320帧中的烧录文本和潜在的受保护健康信息数据。成功去标识化的示例见图 7。所有模型也都未能处理同一超声序列及其全部10帧。该序列的示例帧见图 8。因此，这些模型的去标识化准确率达到96.97%，在330帧中正确分割了320帧。观察到的像素误差均来自心电图等动态组件。因此，它们不太可能属于受保护健康信息，但在训练深度学习模型时仍可能成为潜在的偏差来源。

3 讨论

本文提出的四步模型优化方法展示了一种简单且经济高效的方式，可提高分割模型准确率。该方法不仅适用于高质量图像，还提升了在视觉检查中被计算机视觉算法判定为失败的图像上的性能，并取得了良好的准确率。除第三个凸包折次外，所有细化模型折次的平均性能均优于基础模型。这

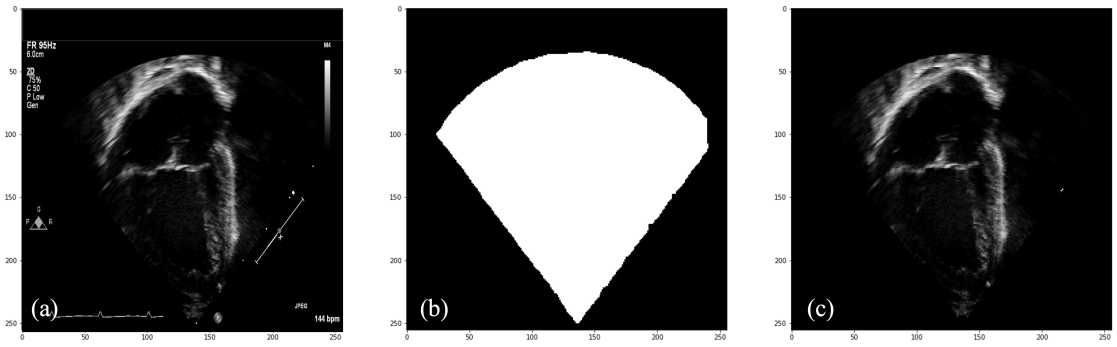

图 7 成功去标识化示例（第 2 次细化）：(a)：包含多个烧录文字位置的原始心脏输入帧，(b)：模型生成的分割掩膜，(c)：去标识化后的掩膜图像。

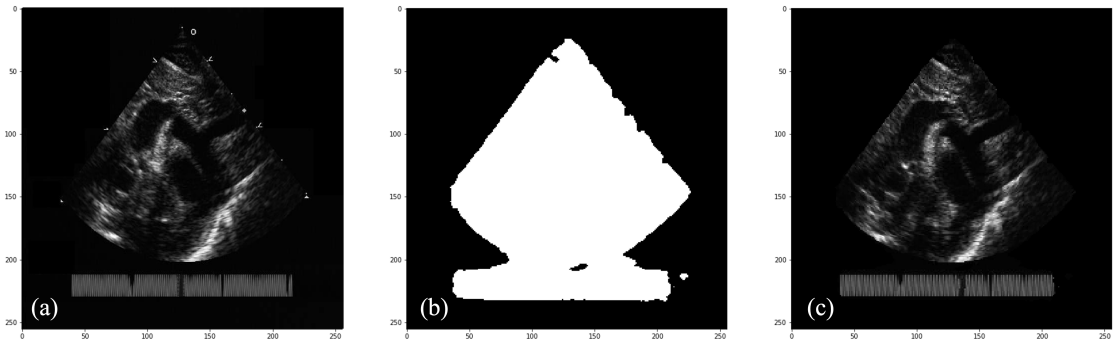

图 8 去标识化失败示例（第 2 次细化）：(a)：包含杂乱移动心电图线的原始心脏输入帧，(b)：模型生成的分割掩膜，(c)：包含心电图线的掩膜图像。

表明，当用户受限于准确率较低的输入训练掩膜时，该方法最为有效。计算机视觉算法的初始性能越好，该方法的边际收益就越小。所提出的多次运行深度学习实验并进行统计计算，以改进实验评估的理念，证明是一种有价值的工具。在所展示的应用案例中，除填充阈值的第三折外，所有性能差异在统计上均显著。该折次与其他折次存在显著差异，可能是因为用于细化的图像数量较少，仅有 22 张，因此一个质量不佳的样本就会对性能产生显著的负面影响。各折次的结果可能会更加相似

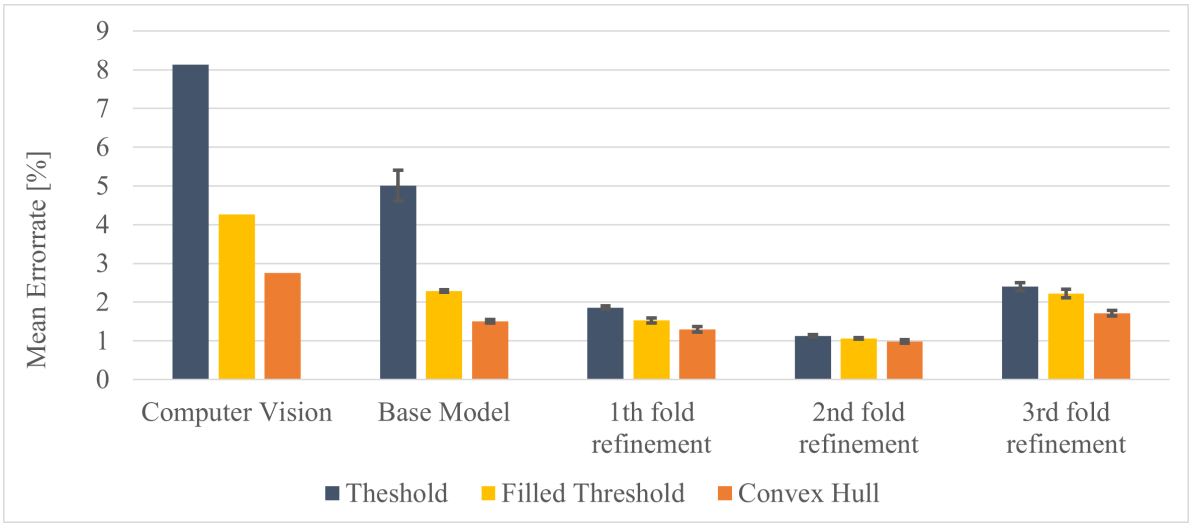

图9 计算机视觉算法、基础模型和细化模型在三种掩膜类型上的平均错误率比较。

或许可以通过增加样本数量来实现。为展示所提出的精炼过程在减少开发复杂计算机视觉算法的必要性方面的有效性，对三种掩膜类型进行了比较。平均错误率（图9）表明，细化模型的性能比之前更为接近。在第二次数据划分中，精炼后的阈值模型与精炼后的凸包模型之间的差异，从最初计算机视觉算法性能差异的5.37%缩小至仅0.15%。这表明，采用模型精炼时，即使是阈值掩膜这样简单的初始训练掩膜，也能取得与凸包掩膜等更复杂输入相近的结果。

由于使用了相同数量的 22 张精炼图像，不同精炼折之间的性能再次出现差异。因此，对于所有模型而言，其中一折中若存在一个错误标签，就会对性能产生显著的负面影响。这也与细化模型的整体结果相符：第三次数据划分在所有掩膜类型上的表现都是最差的。使用更大的数据集来细化基础模型或许可以减小这种影响。不过，所有折中细化模型之间的差异，远小于基础模型最初的差异，或

计算机视觉模型之间的差异，因此证明了所提出的精炼过程带来的巨大改进。

在评估细化模型去除超声图像中的文本、偏置基准标记以及潜在受保护健康信息背景数据的能力时，该方法取得了出色的结果。所有模型都能很好地识别图像中超声锥形区域的位置，彼此之间的差异仅在于对锥体精确形状的匹配准确率。因此，几乎所有模型都能够将回波数据与背景分离。所有模型在全部十帧中都未能对同一测试超声序列进行分割，该序列包含动态心电图。由于测试集规模较小，仅包含 33 个图像序列，因此所有模型的准确率都明显下降。因此，未来的工作应分析：如果增加包含动态背景物体的训练样本，并在规模更大的测试集上评估模型，准确率是否还能进一步提高。不过，所提出的去标识化方法仅适用于能够移除所有背景数据的应用案例；若将非受保护健康信息保留在像素数据中，既没有额外价值，也会带来模型偏差风险。对于需要在图像中保留文本数据的应用，仍需采用更有针对性的方法。

4 结论

结果表明，先在低质量自动分割结果上训练基准模型，再用少量手动分割数据对模型进行细化，这种细化方法能够提升模型性能。通过多次实验运行计算p值，我们证明了模型性能提升具有统计显著性。实验显示，改进主要体现在质量较差的初始输入掩膜上，而对于原本表现良好的输入掩膜，性能提升则逐渐减小。在我们的实验中，不同质量训练掩膜之间的差异

通过采用细化方法得以缩小。这表明，该方法只需利用一个小型准确数据集来细化模型性能，便可减少对大规模高准确率分割掩膜数据集的需求。在我们的实验中，各实验折之间仍存在差异，但这可能是由于用于细化模型的真实标注图像数量较少，只有22张。我们证明，即使初始计算机视觉算法失效，深度学习仍能取得良好效果。所提出的方法证明了其充分发挥深度学习模型优势的能力。这一理念适用于许多单锥超声分割任务；在这些任务中，标注真实标注数据成本高昂，而人工评估自动生成的标签则可以很快完成。对于这类任务，本文展示的实验还说明了一种简单且稳健的方法，可去除静态文本和偏置基准标记，包括其中可能包含的受保护健康信息数据。通过多次运行实验来开展统计分析、验证性能提升结果的方法，是一种实用的验证手段。如果计算成本的增加并非关键限制因素，这种方法便能以更充分的信息为依据比较深度学习实验。

披露声明

作者声明不存在任何财务或其他方面的利益冲突。

致谢

我们谨此感谢 Robert Howe 教授提供的宝贵讨论以及对计算设施的使用权限。

参考文献

1 A. Krizhevsky、I. Sutskever 和 G. E. Hinton，“使用深度卷积神经网络进行 ImageNet 分类”，载于神经信息处理系统进展，F. Pereira、C. Burges、L. Bottou、等编，**25**，Curran Associates, Inc.（2012）。

2 S. Liu 和 W. Deng，“基于小规模训练样本的超深卷积神经网络图像分类”，载于 *2015* 年第 *3* 届 *IAPR* 亚洲模式识别会议（*ACPR*），730–734（2015）。

3 S. Han、H. Mao 和 W. J. Dally，“深度压缩：通过剪枝、训练后量化和霍夫曼编码压缩深度神经网络”，国际学习表征会议（*ICLR*）**4**（2016）。

4 J. Frankle 和 M. Carbin，“彩票假设：寻找稀疏可训练神经网络”，载于第 *7* 届国际学习表征会议（*ICLR 2019*），*2019* 年 *5* 月 *6* 日至 *9* 日，美国路易斯安那州新奥尔良，OpenReview.net（2019）。

5 K. He、X. Zhang、S. Ren、等，“用于图像识别的深度残差学习”，载于*IEEE* 计算机视觉与模式识别会议论文集，770–778（2016）。

6 J. Deng, W. Dong, R. Socher，等，“ImageNet：大型分层图像数据库”，载于 *2009* 年 *IEEE* 计算机视觉与模式识别会议，248–255，IEEE（2009）。

7 D. Choi, A. Passos, C. J. Shallue，等，“通过数据回响加速神经网络训练”，*CoRR* **abs/1907.05550**（2019）。

8 H. Harvey 和 B. Glocker，放射学机器学习任务医学影像数据准备的标准化方法：机遇、应用与风险，61–72。Springer，Cham（2019）。 [doi:10.1007/978-3-319-94878-2 6]。

9 O. Russakovsky、J. Deng、H. Su，等，“ImageNet 大规模视觉识别挑战赛”，国际计算机视觉杂志 *(IJCV)***115**(3)，211–252（2015）。[doi:10.1007/s11263-015-0816-y]。

10 R. Barth、J. IJsselmuiden、J. Hemming，等，“用于语义植物部位分割的卷积神经网络合成自举”，《农业中的计算机与电子学》**161**，291–304（2019）。 [doi:10.1016/j.compag.2017.11.040]。

11 O. Ronneberger、P. Fischer 和 T. Brox，“U-Net：用于生物医学图像分割的卷积神经网络”（2015）。 [doi:10.1007/978-3-319-24574-4 28]。

12 D. P. Kingma 和 J. Ba，“Adam：一种随机优化方法”，载于第三届国际学习表征会议（*ICLR 2015*），美国加利福尼亚州圣迭戈，*2015*年*5*月*7*日至*9*日，会议论文集，Y. Bengio 和 Y. LeCun 编（2015）。

13 Student,《均值的可能误差》，《生物计量学》**6**(1), 1–25（1908年）。

Manuel Zahn 目前领导 VideaHealth 的机器学习团队。在这项工作期间，他曾是哈佛仿生机器人实验室的研究人员。他于 2015 年和 2019 年分别获得慕尼黑工业大学电气工程和信息技术专业的理学学士与理学硕士学位。2017 年，他曾在斯坦福机器人实验室开展水下人机交互研究。他目前的研究兴趣包括机器人学、计算机视觉和深度学习。

Douglas Perrin 是哈佛医学院外科学讲师、波士顿儿童医院高级研究科学家，以及哈佛工程与应用科学学院附属研究员。他于1995年、1999年和2002年分别在明尼苏达大学获得计算机科学理学学士、理学硕士和哲学博士学位。

他在计算机视觉、机器人学、视觉伺服、触觉技术、超声成像与手术引导，以及基于医学成像的儿科心脏病理机器学习等领域发表过论文。

图目录

1 (a)：包含非回波杂波的输入图像，其中可能含有受保护健康信息；(b)：仅包含超声图像数据的二值掩膜；(c)：应用掩膜后，不含任何潜在偏差和受保护健康信息的输入图像。2 数据集中不同心跳序列的若干示例帧，展示了旋转、尺度和结构方面的变化。3 完整的模型精炼方法。4 创建训练掩膜的流程：(a)：n 个原始心脏帧；(b)：n-1 个差分图像的均值；(c)：通过自适应阈值化生成二值掩膜；(d)：应用泛洪填充算法后的掩膜；(e)：应用凸包算法后的掩膜

5 在“代表性测试集”上评估的基础模型每十个训练轮次的平均训练准确率和方差。上排：阈值模型；下排左侧：填充阈值模型；下排右侧：凸包模型

6 基础模型和三个细化划分在代表性测试集上每十个训练轮次评估一次的平均训练准确率和方差。上排：阈值模型；下排左侧：填充阈值模型；下排右侧：凸包模型

7 成功去标识化的示例（第 2 次细化）：(a)：包含多个烧录文字位置的原始心脏输入帧，(b)：模型生成的分割掩膜，(c)：去标识化后的掩膜图像。8 去标识化失败的示例（第 2 次细化）：(a)：包含杂乱移动心电图线的原始心脏输入帧，(b)：模型生成的分割掩膜，(c)：包含心电图线的掩膜图像。9 计算机视觉算法、基础模型和细化模型在三种掩膜类型上的平均错误率比较。

表格目录

1 平均模型准确率（n=6）和 95% 置信区间 2 不同模型的统计显著性评估。3 “Bad Mask 测试集 ” 上的平均模型准确率（ n=6 ）和 95% 置信区间 4 不同模型的统计显著性评估。