HyperSORT：使用超网络的自组织鲁棒训练

Samuel Joutard[1⋆], Marijn Stollenga[1⋆], Marc Balle Sanchez[1,2], Mohammad Farid Azampour[2], and Raphael Prevost[1]

[1] ImFusion, Munich, Germany
[2] Computer Aided Medical Procedures, Technische Universität München, Germany

摘要 医学成像数据集通常包含从错误标签到不一致标注风格的各种异质偏差。这种偏差可能会对深度分割网络的性能产生负面影响。然而，识别和描述这些偏差是一个特别繁琐且具有挑战性的任务。在这篇论文中，我们介绍了 HyperSORT 框架，该框架使用超网络从表示图像和注释变化的潜在向量来预测 UNets 的参数。超网络参数以及训练集中的每个数据样本对应的潜在向量集合是联合学习的。因此，与其优化一个单独的神经网络以拟合数据集，HyperSORT 学习了 UNet 参数的一个复杂分布，在这个分布中低密度区域可以捕捉噪声特定模式，而较大的模态则能够稳健地对器官进行差异但有意义的分割。我们在两个 3D 腹部 CT 公开数据集上验证了我们的方法：首先是 AMOS 数据集的人工扰动版本，以及 TotalSegmentator，一个包含真实未知偏差和错误的大规模数据集。实验表明，HyperSORT 创建了数据集的结构化映射，允许识别相关的系统偏差和错误样本。潜在空间聚类产生的 UNet 参数执行分割任务时符合底层“学习”的系统偏差。代码和我们对 TotalSegmentator 数据集的分析已经公开：https://github.com/ImFusionGmbH/HyperSORT

Keywords: 超网络 · 强健训练 · 自组织。

1 介绍

深度学习解决方案在医学图像分析中的开发需要对训练数据及其标注 [25] 进行彻底审查。确实，诸如错误标注或采集错误等数据不规则情况可能会扰乱训练过程并最终降低最终算法的能力 [21]。医学数据整理仍然严重依赖于人工分析 [8]，这使其成为一个特别耗时且容易出错的步骤。

HyperSORT 通过使用一个额外的隐藏变量来建模标注过程，解决了这个问题。因此，这个隐藏变量可以参数化评分者之间的差异或标注错误。超网

⋆ These authors contributed equally to this work.

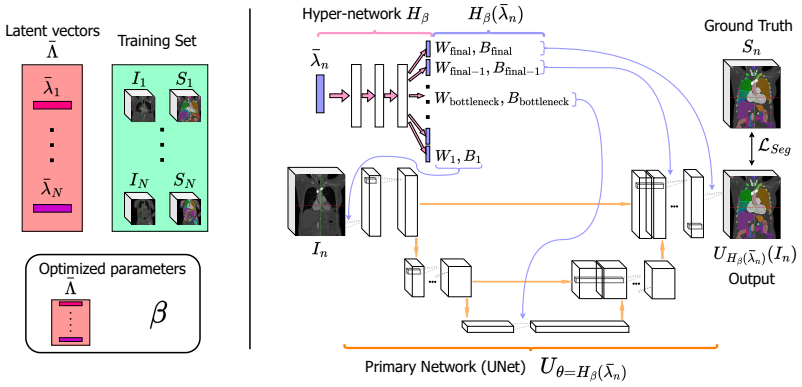

图 1. HyperSORT 概览。超网络 H_β 从特定于训练样本 I_n, S_n 的条件向量 $\bar{\lambda}_n$ 生成 UNet 参数 θ。

络 [10] 根据该变量条件化分割 UNet[20] 的行为。在训练过程中，HyperSORT 联合学习超网络的参数和标注条件隐藏变量的经验分布。所提出方法的概述如图 1 所示。HyperSORT 既提供了稳健训练的分割 UNet 版本，也提供了一个有意义的训练集映射，可用于整理训练集并识别系统性偏差。

我们展示了 HyperSORT 在两个大型 3D 数据集上的性能和可用性。首先作为一个概念验证，其中主要的注释变化模式是已知且受控的，我们在 AMOS 数据集中注入了合成扰动 [14]。其次，作为实际应用场景，我们使用了 TotalSegmentator[23] 的训练集。事实上，这个广受认可的数据集从 V1 到 V2 得到了很大程度上的改进和校正，为异常情况提供了一种伪真实标签。在这些实验中，我们展示了 HyperSORT 生成的性能良好的分割 UNets，并提供了可以解释并用于检测错误标签的有意义的训练集地图。

2 相关工作

数据集质量控制 分割数据整理具有挑战性，因为与分类任务不同，在分类任务中注释要么正确要么错误，而分割掩码可以部分正确和部分错误。虽然已经开发了几种用于回归/分类数据整理的方法（例如 [5]），但关于分割任务的数据整理文献较少。一种典型的方法是依赖重复交叉验证，并使用诸如 Dice 分数等验证指标作为注释质量的代理 [16]。另一种方法可以依赖预训练的质量控制回归器，如最近开发的 Quality Sentinel [6]。尽管这些方法可以

标记一些错误案例，但 HyperSORT 通过提供整个数据集的有意义映射将分析进一步推进。

从噪声标签中学习 为了规避获得金标准标注的挑战，已经开发出了提高模型鲁棒性的方法。当有评估员分层时，可以使用解耦 [26] 或采样重新加权 [18] 。在一般情况下，概率建模允许预测分割分布 [2] 。另外，损失函数 [9,27] 、架构选择 [22,12] 或特定训练策略 [7] 已被证明可以提高模型对错误标签的鲁棒性。关于噪声标签检测和鲁棒性的更多细节，我们建议感兴趣的读者参阅 [24] 以获得更全面的综述。HyperSORT 结合了增强的质量控制和稳健的学习，通过从潜在的嘈杂标签生成高性能网络以及可用于发现错误案例和系统偏差的训练集映射。

超网络 超网络被用作一种方法，根据用户提供的变量来调节主神经网络的行为。在医学成像的背景下，它首次用于动态调整深度可变形配准网络的正则化强度 [11,19]。最近，超网络被用来基于输入图像的空间分辨率条件化一个 3D 分割网络 [15]。超网络还可以通过根据要分割的结构来调节网络，实现从带有异构注释的数据集中协同学习 [3]。所有这些方法都利用了显式条件变量，无论是手工制作还是来自元数据。这里引入的新范式则通过学习和发现训练集中的相关隐式条件来利用超网络。

3 方法

监督分割学习通常假设数据分布 $\mathcal{D}$，从中采样输入/分割对 $(I, S) \sim \mathcal{D}$。一个分割网络，通常是带有参数 U_θ [20] 的 UNetθ，然后被优化以在数据分布下最小化误差度量 $\mathcal{L}_{Seg}(U_\theta(I), S)$。然而，这假设数据标注中的误差是独立同分布（iid）并且围绕实际的“真实标签” [1]。这些假设在医学领域并不总是成立。事实上，可用训练数据的稀缺性和复杂的标注过程（通常依赖于自举、半自动方法 [4] 并展示出高评分者间变异性 [2]）需要一个更为精细的表述。

标签化过程的模型化 相反，我们通过考虑标注过程来更精确地建模数据分布：$\Omega(I, \lambda) \to S$，其中 Ω 是一个未知的确定性预言机函数，而 $\lambda \in \mathcal{R}^n$ 是一个参数化 oracle 注释行为的潜在向量。我们的数据分布明确建模了标签生成过程：$\mathcal{D} = \{I, \Omega(I, \lambda) | I \sim \mathcal{I}; \lambda \sim \Lambda\}$，其中 $\mathcal{I}$ 和 Λ 分别是图像 I 和潜在向量 λ 的分布。λ 向量模型化了标注过程，可以例如表示错误标签或特定的标注

风格来自注释者，正如我们将在第 4 节中更具体地展示的。我们的建模将注释错误分为由 λ 建模的系统性成分和一个中心化的独立同分布加性噪声 [1]，从而放松了学习假设。

超级排序 我们的模型近似于 Oracle 函数 Ω 和注释风格集合 λ 在一个现有的训练集上。首先，我们将一个可训练的潜在向量 $\bar{\lambda}_n$ 关联到每个训练样本 $(I_n, S_n) \in \bar{\mathcal{D}}$ 中，其中 $\bar{\mathcal{D}}$ 是经验数据分布，即训练集。我们考虑由 θ 参数化的已建立的 UNet [20] 架构 U_θ。代替直接优化 θ，我们引入一个超网络 H_β，该超网络由 β [10] 参数化，从潜向量 $\bar{\lambda}$ 预测 UNet 参数 θ。超网络参数 β 和代理潜向量集合 $\bar{\Lambda} = \{\bar{\lambda}_n\}_{n \leq |\bar{\mathcal{D}}|}$ 联合优化为：

$$\min_{\beta, \bar{\Lambda}} \sum_{n=1}^{|\bar{\mathcal{D}}|} \mathcal{L}_{Seg}(U_{H_\beta(\bar{\lambda}_n)}(X_n), S_n) + \mathcal{L}_{reg}(\bar{\lambda}_n) \tag{1}$$

其中，$\mathcal{L}_{Seg}$ 是 Dice 加 CrossEntropy 损失，$\mathcal{L}_{reg}$ 是对潜在向量的 L1 范数正则化项。这个正则化项将潜在向量推向潜在空间的原点，从而使主要标注模式位于零向量 $\overrightarrow{\mathbf{0}}$ 附近。因此，最不寻常的情况通常会远离原点孤立出来并可以被识别。收敛后，超网络 H_β 模仿 OracleΩ，并且学习到的代理潜在向量分布 $\bar{\Lambda}$ 估计注释风格分布 Λ。使用超网络来参数化 Oracle 有两个重要的优势。首先，它允许对注释风格进行低维潜在参数化，在潜在空间中创建训练集的可解释映射。其次，与学习多个不相关的 UNet 参数集合相比，使用超网络已被证明能够实现协同学习 [3]，使不同的注释风格相互受益。

推理 在分割新图像时，我们选择超网络 H_β 将用于调整 UNets 权重的潜在变量。通常情况下，我们会考虑训练过程中形成的潜在向量聚类中心，这些中心对应于不同的标注风格。给定正则化项 $\mathcal{L}_{reg}$，一个典型的选择是使用 $\lambda = \overrightarrow{\mathbf{0}}$ 作为训练集中主要标注样式的代表。或者，可以由用户动态选择最相关的标注样式来进行分割。此外，给定一组由注释者选定的首选标注风格，HyperSORT 提供了对应的 UNet 参数，这些参数可用于纠正错误标签并生成更好的伪标签。

4 实验与结果

HyperSORT 对网络架构的选择不敏感。由于我们针对分割应用，因此使用了一个标准的 3D UNet 架构，包含 3 个下采样阶段，在最高分辨率

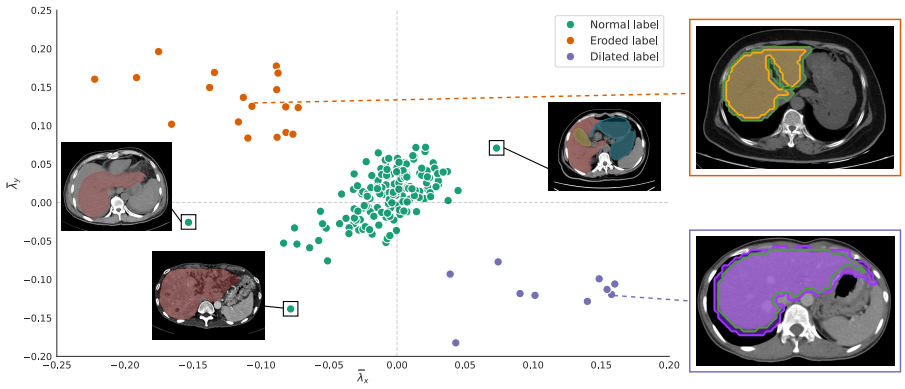

图 2. （左）获得的 AMOS 潜在空间。$\overrightarrow{\mathbf{0}}$ 聚类中最极端的情况是具有挑战性的案例（从左到右）：不完整的肝脏分割、肝组织异质性以及异常的腹部解剖结构，其中胆囊和胃作为参考被显示出来。（右）使用 $\overrightarrow{\mathbf{0}}$ 潜在变量对来自侵蚀和膨胀聚类的图像进行推断。

处有 16 个通道，并且在每个分辨率阶段都有 3 个卷积层（ReLU 激活函数，之后是实例归一化）。我们使用了二维潜在向量来促进结果的可视化和分析。超网络仅由一个全连接网络组成，该网络包含三个隐藏层，每层大小为 50（ReLU 激活函数）。最终 UNet 参数预测后跟随一个自定义激活功能 $x :\longrightarrow tanh(x) * 5$，限制预测参数的范数。这在实验上稳定了训练。所有参数均使用 Adam 优化器进行训练，初始学习率为 10^{-4}，直到收敛。更多详情可以在我们的公共仓库 [1] 中找到。

4.1 概念验证使用合成标签扰动

作为第一个概念验证，我们创建了一个多评分员场景的粗略近似值，其中一些评分员在器官边界方面比其他人更保守。我们从 AMOS 训练数据集 [14] 中导出了一个包含 200 次 CT 扫描的肝脏分割数据集。我们将 $\sim$ 数据集中的 15% 进行扰动，通过对 $\sim$ 的 7.5% 扫描执行 3、4 或 5 轮侵蚀操作，并对另一部分 $\sim$ 的 7.5% 执行扩张，剩余的 85% 标签未被扰动。学习到的潜在向量分布 $\bar{\Lambda} \subset \mathbb{R}^2$ 如图 2 所示。我们观察到，方向 $[1.0, -1.0]$ 捕捉到了肝脏边界的紧致程度。此外，在潜在空间中，合成标签样式（正常、侵蚀和膨胀）被清晰地分开并以有意义的方式排序。侵蚀和膨胀聚类沿该方向更加扩散，因为它们包含了每次形态学操作应用次数的变化性。我们还看到中心

[1] https://github.com/ImFusionGmbH/HyperSORT

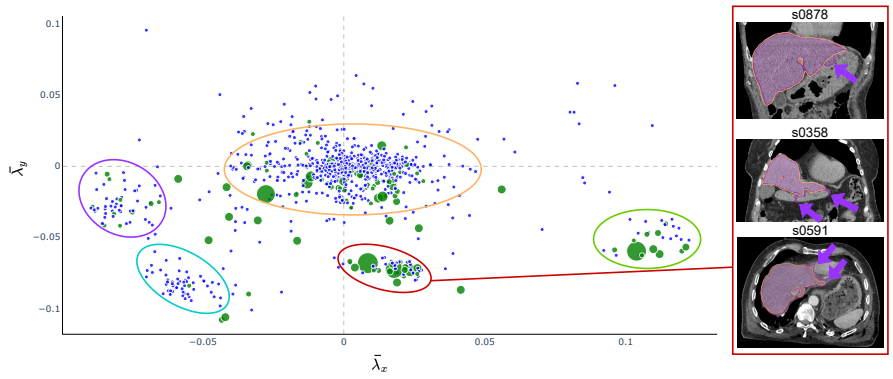

图 3. （左）获得的 TS 潜在空间。蓝色和绿色点分别对应从 V1 到 V2 的未校正和已校正案例。绿色点的半径与其修改的体素数量成正比。用彩色椭圆高亮显示了 5 个视觉聚类。（右）来自■聚类的 3 个案例的切片，在 V1 到 V2 之间未被修改。紫色箭头突出显示错误的肝脏标记。展示了来自■聚类 UNet 的校正预测。

聚类利用了另一个正交的方向来捕捉一些额外的次要变化性。例如，如图 2 所示，远离该聚类的三个最远案例都是具有挑战性的。最后，如图 2 右侧所示，来自优选聚类的 UNet 参数可以用于校正训练集中存在的错误注释，使得 HyperSORT 成为引导场景下特别方便的工具。

4.2 应用于 TotalSegmentator 数据集

TotalSegmentator (TS) 数据集 [23] 是一个包含来自不同机构、扫描仪和协议的 1204 个标注 CT 扫描图像的综合集合。相应的标签图包含了超过 100 种解剖结构。其庞大的规模使其在研究界非常受欢迎，并成为许多论文的基础。然而，由于其标注过程（迭代学习，通过手动细化现有模型的预测），这些标签图有时可能包含伪影和过度/不足分割的情况。因此，最近发布了该数据集的第二个版本（TS-V2），旨在修复其中的一些错误。这使其成为一个适合我们方法测试的平台，因为我们能够将 TS-V1 作为“有缺陷”的数据集，并使用 TS-V2 作为代理真实值。从 V1 到 V2，$\sim 50\%$ 肝脏标注得到了纠正，$\sim 20\%$ 则需要进行重大调整（>10000 体素更改）。该实验使我们能够证明我们的两个主要论点：

聚类捕获注释“风格”并生成稳健的网络 在这个实际用例中，图 3 中所展示的五个聚类的意义更为微妙。然而，我们在这里展示了它们都能生成具

有显著“标注风格”差异的可使用的 UNet。我们将每个这些聚类中心处获得的五组 UNet 参数集进行了考虑。相比之下，我们训练了五个与主网络架构相同的随机初始化的 UNets。我们也将其与使用 nnUNet 框架 [13] 训练的 TotalSegmentator 模型 [23] 进行比较，并用一个使用 T-loss [9] 训练的“鲁棒 UNet”作为性能参考。我们在 CT-1K 数据集子任务 2[17] 上评估这些模型，这是一个包含 361 个多样化的腹部 CT 扫描的大数据集，并且与 TS 没有重叠。性能结果见表 1。对于 5 个随机初始化的模型和从 HyperSORT 聚类中心得出的 5 个模型，我们还报告了“五个模型中最佳”的性能，模拟了一个带有“人工循环”推理场景。我们观察到所有由 HyperSORT 生成的 UNets 在大型测试集上都表现出具有竞争力的表现。特别地，TS 模型（nnUNet），被视为最先进的模型，在我们的所有模型范围内，验证了我们的实验设置。鲁棒模型与标准 UNets 表现相似，这是因为 TS 数据集的规模和整体良好的质量。我们还注意到即使包含有限数量样本的小型聚类也实现了良好的泛化性能，我们认为这是由于超网络中协同学习能力在 [3] 中所强调的内容。

最重要的是，我们注意到 HyperSORT 通过提供五种不同的 UNet 来分割肝脏，从而更好地探索了解空间。事实上，在病例之间每个 UNet 预测的 Dice 标准差平均值在 HyperSORT UNets 中（0.5）是五种随机初始化的 UNets 中的两倍大（0.2）（$p_{value} \leq 10^{-5}$）。这种对解空间的精细探索系统地提高了“人机交互”推理场景下的预测效果（$p_{value} \leq 10^{-5}$）。由于获得的解决方案在保持有意义的同时更加分化，HyperSORT 学习到的潜在空间左侧的注释风格（例如 ■ 簇）必须更好地对应于 CT-1k 数据集标签的注释风格。

除了这种定量评估之外，我们在图 3 中观察到红色聚类包含了一些从 V1 修正到 V2 的案例。该图还突出了该聚类中的几个未被纠正的样本，并展示了错误的标注。这表明这个聚类捕捉到了一种特定形式的系统性注释错误，解释了那个聚类的 UNet 在 CT-1K 数据集上表现较差的原因。这也证实了 HyperSORT 作为一个工具的实际价值，它能够提供对数据集的丰富而有意义的分析以及强大且多样的 UNet 参数。

潜映射可用于识别错误案例 为了评估 HyperSORT 检测错误案例的能力，我们使用从 V1 到 V2 的肝脏标签的变化作为伪地面实况。只有至少有 1 个体素发生变化的案例才被考虑作为伪地面实况，因为我们确定这些案例是从 V1 到 V2 经过检查的。我们比较了四种不同的预测器用于此实验。首先，Quality Sentinel[6] 作为最近发布的分割标注质量回归器。然后，我们仅在 V1 和 V2 之间未修改的案例上训练一个 UNet（与 HyperSORT 的主

表 1. 模型在 CT-1k 数据集上的性能。颜色对应于从 HyperSORT 潜在聚类中心获得的 UNets。

UNet seed	1	2	3	4	5	Best
Dice (std)	96.4 (1.8)	96.1 (1.5)	96.4 (1.3)	96.5 (1.4)	96.4 (1.5)	96.6 (1.2)
HyperSort UNet						Best
Dice (std)	96.4 (1.3)	96.7 (1.5)	97.1 (1.5)	95.9 (1.4)	96.4 (1.5)	97.2 (1.4)
Additional baselines	时间序列模型			T-损失 UNet		
Dice (std)	97.0 (1.1)			96.5 (1.4)		

要网络相同的架构)。如 [16] 中解释，我们可以使用该模型在训练集的剩余部分上的 Dice 损失作为标签质量的代理。我们将此基线称为 “Test-Dice”。Test-Dice 优于其他预测器，因为训练/测试集划分是利用伪地面实况完成的。请注意，该 UNet 的泛化能力与在整个数据集上训练的 UNet 相当（在 CT-1k 数据集上为 96.4 的测试 Dice 分数)。我们考虑了从 HyperSORT 的训练集映射中得出的两种可能的预测器。根据我们零聚类捕捉规范行为的假设，我们使用代理潜在向量范数 $\{||\bar{\lambda}_n||_2\}_{n\leq|\bar{\mathcal{D}}|}$。此外，作为训练案例的 “隔离” 度量，我们考虑与所有案例的平均距离 $\{\frac{1}{|\bar{\mathcal{D}}|}\sum_m ||\bar{\lambda}_m - \bar{\lambda}_n||_2\}_{n\leq|\bar{\mathcal{D}}|}$。这两个从 HyperSORT 派生的预测器分别与 V1 和 V2 之间修改的体素数量具有 0.2166 和 0.1723 的 Spearman 相关性。另一方面，Quality Sentinel 负分与变化量具有意外的负相关性 (-0.0499)。Test-Dice 尽管具有优势，但也实现了较低的相关性分数 0.1150。因此，这两个从 HyperSORT 派生的特征与从 TS V1 到 V2 应用的修改量更好地相关。此外，我们强调获得的潜在向量图 $\{\bar{\lambda}_n\}_{n\leq|\bar{\mathcal{D}}|}$ 描述了训练集，而不仅仅是错误标签检测，如前所示。这突出了 HyperSORT 识别标签改进候选的能力。

4.3 讨论

我们展示了 HyperSORT 能够捕捉可能影响模型质量的数据集中的异常值和变化。然而，一个待解决的问题是如何区分 “错误标签” 与 “具有挑战性的正确标签”，两者都可能与较大的潜在向量相关联，从而阻止它们在模型分布的主要模式中被表示。另一方面，这也可以指示出数据分布中未充分采样的群体，这些原本会被忽略的部分可以帮助揭示现有数据集中的偏见。关于选择二维的潜在向量，它便于视觉检查，并且在我们的实验中足够用来

捕捉有意义的变化。更高维度的潜在向量可能允许潜在空间欧几里得范数与注释风格变化之间有更均匀的关系，从而有助于聚类解释。评估其他数据集是否需要更高的潜在空间维数留待未来的工作。最后，虽然为了简洁起见我们在本文中重点关注了 CT 中的肝脏分割问题，但 HyperSORT 可以应用于任何分割任务，包括具有挑战性的结构如肠道或多分类问题。这类复杂任务更有可能表现出样本数据内部叠加的系统性偏见，使得使用常规聚类方法识别这些偏见更加困难。我们的公开仓库使将 HyperSORT 扩展到任何架构变得特别直接，并展示了在一系列知名公共数据集上获得的潜在空间。我们希望这有助于进一步整理和改进这些数据集。

5 结论

在这篇论文中，我们介绍了 HyperSORT，它以一种新颖的方式利用超网络对训练集进行精细分层，并在生成性能稳健的训练网络的同时帮助识别错误案例和系统偏差。如我们的实验所示，HyperSORT 同时充当结构化整理和校正工具，可以在使用大型数据集训练新模型时系统地使用。

Disclosure of Interests. 作者声明与本文内容相关的不存在任何利益冲突。

参考文献

1. Bach, F.: Learning Theory from First Principles. Adaptive Computation and Machine Learning series, MIT Press (2024)
2. Baumgartner, C.F., Tezcan, K.C., Chaitanya, K., Hötker, A.M., Muehlematter, U.J., Schawkat, K., Becker, A.S., Donati, O., Konukoglu, E.: Phiseg: Capturing uncertainty in medical image segmentation. In: Shen, D., Liu, T., Peters, T.M., Staib, L.H., Essert, C., Zhou, S., Yap, P.T., Khan, A. (eds.) Medical Image Computing and Computer Assisted Intervention – MICCAI 2019. pp. 119–127. Springer International Publishing, Cham (2019)
3. Billot, B., Dey, N., Turk, E.A., Grant, E., Golland, P.: Network conditioning for synergistic learning on partial annotations. In: Medical Imaging with Deep Learning (2024), https://openreview.net/forum?id=sfjgmuvLS7
4. Budd, S., Robinson, E.C., Kainz, B.: A survey on active learning and human-in-the-loop deep learning for medical image analysis. Medical Image Analysis **71**, 102062 (2021). https://doi.org/https://doi.org/10.1016/j.

media.2021.102062, https://www.sciencedirect.com/science/article/pii/S1361841521001080

5. Chen, J., Ramanathan, V., Xu, T., Martel, A.L.: Detecting noisy labels with repeated cross-validations **LNCS 15010** (October 2024)

6. Chen, Y., Zhou, Z., Yuille, A.L.: Quality sentinel: Estimating label quality and errors in medical segmentation datasets. CoRR **abs/2406.00327** (2024), https://doi.org/10.48550/arXiv.2406.00327

7. Dong, W., Du, B., Xu, Y.: Shape-intensity knowledge distillation for robust medical image segmentation. Frontiers of Computer Science **19**(9), 199705 (Jan 2025). https://doi.org/10.1007/s11704-024-40462-2, https://doi.org/10.1007/s11704-024-40462-2

8. Galbusera, F., Cina, A.: Image annotation and curation in radiology: an overview for machine learning practitioners. European Radiology Experimental **8**(1), 11 (Feb 2024). https://doi.org/10.1186/s41747-023-00408-y, https://doi.org/10.1186/s41747-023-00408-y

9. Gonzalez-Jimenez, A., Lionetti, S., Gottfrois, P., Gröger, F., Pouly, M., Navarini, A.A.: Robust t-loss forămedical image segmentation. In: Greenspan, H., Madabhushi, A., Mousavi, P., Salcudean, S., Duncan, J., Syeda-Mahmood, T., Taylor, R. (eds.) Medical Image Computing and Computer Assisted Intervention – MICCAI 2023. pp. 714–724. Springer Nature Switzerland, Cham (2023)

10. Ha, D., Dai, A.M., Le, Q.V.: Hypernetworks. In: International Conference on Learning Representations (2017), https://openreview.net/forum?id=rkpACe1lx

11. Hoopes, A., Hoffmann, M., Greve, D.N., Fischl, B., Guttag, J., Dalca, A.: Learning the effect of registration hyperparameters with hypermorph. Machine Learning for Biomedical Imaging **1**, 1–30 (2022). https://doi.org/10.59275/j.melba.2022-74f1

12. Iqbal, S., Khan, T.M., Naqvi, S.S., Naveed, A., Meijering, E.: Tbconvl-net: A hybrid deep learning architecture for robust medical image segmentation. Pattern Recognition **158**, 111028 (2025). https://doi.org/https://doi.org/10.1016/j.patcog.2024.111028, https://www.sciencedirect.com/science/article/pii/S0031320324007799

13. Isensee, F., Jaeger, P.F., Kohl, S.A.A., Petersen, J., Maier-Hein, K.H.: nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. Nature Methods **18**(2), 203–211 (Feb 2021). https://doi.org/10.1038/s41592-020-01008-z, https://doi.org/10.1038/s41592-020-01008-z

14. Ji, Y., Bai, H., Yang, J., Ge, C., Zhu, Y., Zhang, R., Li, Z., Zhang, L., Ma, W., Wan, X., et al.: Amos: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation. arXiv preprint arXiv:2206.08023 (2022)

15. Joutard, S., Pietsch, M., Prevost, R.: HyperSpace: Hypernetworks for spacing-adaptive image segmentation . In: proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2024. vol. LNCS 15009. Springer Nature Switzerland (October 2024)

16. Lad, V., Mueller, J.: Estimating label quality and errors in semantic segmentation data via any model. arXiv preprint arXiv:2307.05080 (2023)

17. Ma, J., Zhang, Y., Gu, S., Zhang, Y., Zhu, C., Wang, Q., Liu, X., An, X., Ge, C., Cao, S., Zhang, Q., Liu, S., Wang, Y., Li, Y., Wang, C., He, J., Yang, X.: Abdomenct-1k: Is abdominal organ segmentation a solved problem? IEEE Transactions on Pattern Analysis and Machine Intelligence **44**, 6695–6714 (2020), https://api.semanticscholar.org/CorpusID:225094385

18. Mirikharaji, Z., Yan, Y., Hamarneh, G.: Learning to segment skin lesions from noisy annotations. CoRR **abs/1906.03815** (2019), http://arxiv.org/abs/1906.03815

19. Mok, T.C.W., Chung, A.C.S.: Conditional deformable image registration with convolutional neural network. In: de Bruijne, M., Cattin, P.C., Cotin, S., Padoy, N., Speidel, S., Zheng, Y., Essert, C. (eds.) MICCAI 2021. pp. 35–45. Springer International Publishing, Cham (2021)

20. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015 (2015)

21. Sylolypavan, A., Sleeman, D., Wu, H., Sim, M.: The impact of inconsistent human annotations on ai driven clinical decision making. npj Digital Medicine **6**(1), 26 (Feb 2023). https://doi.org/10.1038/s41746-023-00773-3, https://doi.org/10.1038/s41746-023-00773-3

22. erban Vădineanu, Pelt, D., Dzyubachyk, O., Batenburg, J.: An analysis of the impact of annotation errors on the accuracy of deep learning for cell segmentation. In: Medical Imaging with Deep Learning (2022), https://openreview.net/forum?id=C4B46ZS7MSB

23. Wasserthal, J., Breit, H.C., Meyer, M.T., Pradella, M., Hinck, D., Sauter, A.W., Heye, T., Boll, D.T., Cyriac, J., Yang, S., Bach, M., Segeroth, M.: Totalsegmentator: Robust segmentation of 104 anatomic structures in ct images. Radiology: Artificial Intelligence **5**(5), e230024 (2023). https://doi.org/10.1148/ryai.230024, https://doi.org/10.1148/ryai.230024

24. Wei, Y., Deng, Y., Sun, C., Lin, M., Jiang, H., Peng, Y.: Deep learning with noisy labels in medical prediction problems: a scoping review. Journal of the American Medical Informatics Association **31**(7), 1596–1607 (05 2024). https://doi.org/10.1093/jamia/ocae108, https://doi.org/10.1093/jamia/ocae108

25. Willemink, M.J., Koszek, W.A., Hardell, C., Wu, J., Fleischmann, D., Harvey, H., Folio, L.R., Summers, R.M., Rubin, D.L., Lungren, M.P.: Preparing medical imaging data for machine learning. Radiology **295**(1), 4–15 (2020). https://doi.org/10.1148/radiol.2020192224, https://doi.org/10.1148/radiol.2020192224, pMID: 32068507

26. Zhang, L., Tanno, R., Xu, M.C., Jacob, J., Ciccarelli, O., Barkhof, F., C. Alexander, D.: Disentangling human error from the ground truth in segmentation of medical images. NeurIPS (2020)

27. Zhang, Z., Sabuncu, M.R.: Generalized cross entropy loss for training deep neural networks with noisy labels. In: Proceedings of the 32nd International Conference on Neural Information Processing Systems. p. 8792 – 8802. NIPS'18, Curran Associates Inc., Red Hook, NY, USA (2018)