

电 子 科 技 大 学

UNIVERSITY OF ELECTRONIC SCIENCE AND TECHNOLOGY OF CHINA

博士学位论文

DOCTORAL DISSERTATION



论文题目 基于有限标注数据的医学图像分割

学科专业 机械工程

学 号 202011040814

作者姓名 罗祥德

指导老师 张少霆 教授

学 院 机械与电气工程学院

分类号 TP391.41 密级 公开
UDC 注 1 004.9

学 位 论 文

基于有限标注数据的医学图像分割

(题名和副题名)

罗祥德

(作者姓名)

指导教师 张少霆 教 授
电子科技大学 成 都

(姓名、职称、单位名称)

申请学位级别 博士 学科专业 机械工程
提交论文日期 2024 年 4 月 30 日 论文答辩日期 2024 年 5 月 28 日
学位授予单位和日期 电子科技大学 2024 年 6 月
答辩委员会主席 李康
评阅人 王国泰、朱晓峰、523178370、523178404、530282805

注 1: 注明《国际十进分类法 UDC》的类号。

Medical Image Segmentation Using a Limited Amount of Labeled Data

A Doctoral Dissertation Submitted to
University of Electronic Science and Technology of China

Discipline:	Mechanical Engineering
Student ID:	202011040814
Author:	Luo Xiangde
Supervisor:	Zhang Shaoting Professor
School:	School of Mechanical and Electrical Engineering

独创性声明

本人声明所呈交的学位论文是本人在导师指导下进行的研究工作及取得的研究成果。据我所知，除了文中特别加以标注和致谢的地方外，论文中不包含其他人已经发表或撰写过的研究成果，也不包含为获得电子科技大学或其它教育机构的学位或证书而使用过的材料。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示谢意。

作者签名： 罗祥德

日期： 2024 年 6 月 19 日

论文使用授权

本学位论文作者完全了解电子科技大学有关保留、使用学位论文的规定，同意学校有权保留并向国家有关部门或机构送交论文的复印件和数字文档，允许论文被查阅。本人授权电子科技大学可以将学位论文的全部或部分内容编入有关数据库进行检索及下载，可以采用影印、扫描等复制手段保存、汇编学位论文。

（涉密的学位论文须按照国家及学校相关规定管理，在解密后适用于本授权。）

作者签名： 罗祥德

导师签名： 张少军

日期： 2024 年 6 月 19 日

摘 要

图像分割是计算机辅助诊断系统中一项基础且关键的任务，特别是在图像引导的手术规划和放疗计划生成等方面。最近，随着深度学习的发展，卷积神经网络在许多自动图像分割任务中取得了可喜的进展。然而，基于全监督学习的卷积神经网络往往需要大量精细标注数据来训练模型以获得良好的分割性能。缺乏大规模高质量的标注数据限制了高性能深度学习模型的开发，进而限制了深度学习模型的临床适用性。因此，如何利用少量标注数据开发高效率、高性能、临床适用的数据标注方法和模型训练策略变得尤为重要。在实践过程中，交互式分割可利用有限标注数据和少量用户交互开发模型对未标注图像进行快速标注。为了减少数据标注成本，弱监督学习展现了其利用稀疏标注（点、线、框或类别信息）来开发高性能模型的潜力，并在多个任务上取得了良好的实验结果。不仅如此，在实际临床场景中存在大量未标注数据，因此半监督学习也可以有效利用少量标注数据和大量未标注数据来开发模型进而缓解深度学习对标注数据的依赖。同时，在数据标注过程中选择信息量大的样本进行标注能最大程度提高模型的分割性能，因此利用主动学习来选取少量高信息量样本进行标注和模型训练也是减少数据标注成本的有效手段。本论文的主要工作和创新如下：

1. 针对医学图像分割标注过程中效率低和成本高等问题，本文提出了一种基于测地线距离变换的可泛化交互式分割标注。该方法从用户交互流程、交互信息编码和分割结果修正三个方面结合深度学习算法开发了一种新的交互式分割流程。首先根据医学图像形状和灰度分布特点，本文创新性地提出了内部边界点作为用户交互，减少用户交互量的同时更好地表示分割目标的解剖结构；同时提出了指数化测地线距离来编码用户交互，突出分割目标的同时减少超参数依赖，最后提出了基于信息融合的图割算法用于结果修正。多中心多任务分割数据集上的实验表明，该方法可以以极少的交互量实现准确的图像分割，三维图像标注的平均精度为 88.71%，与主流的标注软件 ITK-SNAP 相比精度提高了 6%，且标注效率提高了近五倍。此外，该方法还可以在不重新训练和微调模型的情况下用于未见过的模态或者目标类别进行交互式标注。

2. 针对有限的稀疏标注监督信号不足的问题，本文提出了一种基于动态混合伪标签的涂鸦监督医学图像分割方法。该方法提出了一种弱监督训练策略来促进深度学习模型从稀疏的涂鸦标注中学习并实现准确的分割。首先提出了双分支网络用于生成多样性预测以克服伪标签不易更新的内在缺陷。之后，基于双分支网络

的多样性预测，本文进一步提出了动态混合伪标签方法来生成高质量的伪标签用于模型训练。该方法在心脏结构分割和腹部多器官分割数据集上取得了准确的分割结果，与单纯地从标注像素中学习相比，本方法的分割结果分别提高了 19.54% 和 11.33%，证明该方法可以提高深度学习模型从涂鸦标注中学习的能力。

3. 针对实际临床场景存在的大量无标注数据和有限的标注数据问题，本文提出了一种基于双任务一致性的半监督医学图像分割方法。该方法利用图像分割任务的不同表示方法设计了双任务预测网络，同时通过任务转换函数把两个不同任务映射到同一表示形式。进一步为了迫使两个任务预测结果相似，构建了跨任务的一致性约束。左心房分割和胰腺分割数据集上的实验结果表明，该方法只需要 20% 左右的标注数据就可以获得与全标注相当的分割结果，证明了双任务一致性可以鼓励模型从无标注数据中学习进而降低模型对标注数据的依赖。

4. 针对多中心临床应用过程中模型泛化性有限等问题，本文提出了一种基于主动学习的跨域医学图像分割方法及进行了多中心临床验证。本方法针对现有深度学习方法依赖分布一致性数据和未考虑图像信息量差异问题，首先提出了一个新的无源主动域适应场景；同时为了降低标注成本，本方法利用主动学习和源域模型从目标域中选择信息量高的部分样本进行标注和模型微调。为了验证该方法的有效性，本论文构建了多中心和多标注者鼻咽癌靶区勾画数据集，实验结果表明该方法可以在有限标注数据下实现准确的跨域分割，如主动选择 20% 左右数据进行标注即可取得与全标注相当的分割性能。此外，多中心临床验证表明，该方法生成的预测结果只需要少量医生修正即可临床使用，其中平均修改程度不超过 9%、修改用时少于两分钟，与手工勾画相比每个病人勾画效率提高了近五倍。

本论文针对标注数据缺乏和标注形式多样对高性能医学图像分割模型开发的巨大挑战，围绕基于有限标注数据的医学图像分割问题展开研究，开发了一系列基于深度学习的高效率数据标注方法和模型训练策略，以缓解医学图像分割中模型对大规模标注数据依赖的问题。

关键词：医学图像分割，交互式分割，半监督学习，弱监督学习，主动学习

ABSTRACT

Medical image segmentation is a fundamental and critical task in computer-aided diagnosis systems, especially in applications such as image-guided surgical planning and radiotherapy planning. Deep learning with convolutional neural network (CNN) has recently made remarkable progress in many automated image segmentation tasks. However, fully supervised learning typically demands a substantial amount of finely annotated data to achieve satisfactory segmentation performance. The scarcity of large-scale high-quality annotated data poses a challenge to developing high-performance deep learning models, consequently restricting their clinical applicability. Therefore, it becomes crucial to develop efficient data annotation methods and model training strategies using limited annotated data. As an alternative, interactive segmentation methods can utilize limited annotated data and minimal user interaction to develop models for image annotating rapidly. In addition, weakly supervised learning has also shown its potential to develop high-performance models using sparse annotations (points, lines, bounding boxes, or image-level tags) and has achieved promising experimental results on multiple tasks. Furthermore, since there is an amount of unlabeled data in real clinical scenarios, semi-supervised learning can be used to develop models combining limited labeled data and huge unlabeled data. Moreover, employing active learning to select a limited number of informative samples for annotation and model fine-tuning is also an effective means to address the high cost of data annotation. The main contributions and innovations of this dissertation are summarized as follows:

1. To alleviate the challenges of low efficiency and high cost during the medical image segmentation annotation, this dissertation proposed a new method for generalizable interactive segmentation and annotation based on geodesic distance transformation. This method develops a novel interactive segmentation workflow combining deep learning algorithms from three aspects: user interaction pipeline, interaction encoding, and segmentation refinement. Firstly, considering the characteristics of medical images in shape and intensity distribution, this dissertation proposes using new interior margin points as user interactions to reduce the amount of user interaction and provide more shape information for different types of organs with complex and irregular shapes. Simultaneously, a new context-aware and parameter-free exponential geodesic distance transformation is pro-

posed to encode user interaction. Finally, a novel information fusion followed by a graph cuts algorithm is proposed for segmentation refinement. Experiments on multi-center, multi-task segmentation datasets demonstrate that this method can achieve accurate image segmentation with minimal user interaction, where compared with the widely-used annotation tool ITK-SNAP, the average accuracy of annotating a 3D image is improved by about 6% and the annotation time is reduced by nearly five times. Furthermore, this method can be applied to interactive annotation of unseen modalities or target categories without the need for retraining or fine-tuning the model.

2. In response to the problem of limited supervision signals due to sparse annotations, this dissertation designed a scribble-supervised medical image segmentation method via a dual-branch network and dynamically mixed pseudo labels supervision. This method proposes a novel weakly supervised segmentation algorithm that facilitates deep learning methods to learn from scribble annotations and achieve accurate segmentation results. Specifically, this study first introduces a dual-branch network for generating diverse predictions to overcome the inherent drawback of difficulty in updating pseudo labels. Subsequently, this study further proposed a dynamic mixing pseudo-label method via dynamic combining the dual-branch network predictions to generate high-quality pseudo-labels for model training. The method achieves promising segmentation results on the cardiac structure segmentation and abdominal multi-organ segmentation datasets where compared with learning from annotated pixels this method improves segmentation accuracy by 19.54% and 11.33%, respectively. These encouraging results demonstrated the proposed method can enhance the learning capability of deep learning models from scribble annotations.

3. To alleviate the challenges of labeled data are limited and unlabeled data is huge which widely exists in real clinical scenarios, this dissertation presented a semi-supervised medical image segmentation via dual-task consistency. This method leverages different representations of image segmentation tasks to develop a dual-task segmentation network. Then, two different tasks are mapped to the same representation space via a differentiable task transformation function. Afterwards, a cross-task consistency constraint is constructed to enforce similarity between the predicted results of the two tasks. Experiments on left atrium segmentation and pancreas segmentation datasets showed that the proposed method achieves comparable segmentation results to fully annotated data with only about 20% of the labeled data, suggesting that the proposed method can encourage the model to learn from unlabeled data through dual-task consistency constraints.

4. To tackle the issue of weak model generalization in multi-center clinical applications, this dissertation proposed an active learning based source free domain adaptation method and conducted multi-center clinical assessment. This method alleviates the reliance of existing deep learning methods on distribution-consistent data and the neglect of differences in image information by first proposing a novel unsupervised active domain adaptation scenario. Furthermore, this dissertation utilizes active learning methods and a pre-trained model on the source domain to select the most informative samples from the target domain for annotation and model fine-tuning. To validate the effectiveness of this approach, multi-center and multi-rater datasets for nasopharyngeal carcinoma primary tumor volume delineation are constructed. Experimental results demonstrate that this method achieves accurate cross-domain segmentation with limited annotated data, with segmentation performance comparable to fully annotated data achieved by annotating only about 20% of the data actively selected. Furthermore, multi-center clinical validation shows that the predicted results of this method require minimal doctor correction for clinical use, with an average revision degree not exceeding 9% and revision time less than two minutes, resulting in nearly five times higher segmentation efficiency per patient compared to manual delineation.

This work focuses on the huge challenges posed by the lack of labeled data and the diverse forms of annotations to the development of high-performance medical image segmentation models. It conducts research on the problem of medical image segmentation based on limited labeled data and develops a series of high-efficiency data labeling methods and model training strategies based on deep learning to alleviate the problem of medical image segmentation models' reliance on large-scale labeled data.

Keywords: Medical Image Segmentation, Interactive Segmentation, Weakly-Supervised Learning, Semi-Supervised Learning, Active Learning

目 录

第一章 绪 论	1
1.1 课题背景及研究意义	1
1.2 国内外研究现状	4
1.2.1 图像分割发展现状	4
1.2.2 高效数据标注方法研究现状	6
1.2.3 基于有限标注的学习方法研究现状	7
1.3 主要研究内容	9
1.4 本文组织结构	11
第二章 医学图像分割基础与有限标注学习	14
2.1 医学图像分割基础	14
2.1.1 水平集方法	14
2.1.2 图割法	15
2.1.3 深度学习基础	15
2.1.4 基于深度学习的图像分割	20
2.2 高效标注与有限标注学习	22
2.2.1 交互式分割	22
2.2.2 弱监督学习	24
2.2.3 半监督学习	25
2.2.4 主动学习	26
2.3 有效标注分割数据集	28
2.3.1 多模态脑肿瘤分割 MRI 数据集	28
2.3.2 左心房分割 MRI 数据集	29
2.3.3 腹部多器官 CT 数据集	30
2.3.4 心脏结构分割 MRI 数据集	31
2.3.5 鼻咽癌靶区勾画数据集	31
2.4 本章总结	33
第三章 基于测地线距离变换的可泛化交互式分割标注	36
3.1 引言	36
3.2 算法设计	37
3.2.1 基于内部边界点的用户交互	37

3.2.2	指数化测地线距离	39
3.2.3	基于提示图和模型的初始分割	40
3.2.4	基于初始分割和修正点信息融合的结果修正	40
3.2.5	实现细节	42
3.2.6	对比方法及评价指标	43
3.3	二维医学图像实验验证	44
3.3.1	二维实验数据	44
3.3.2	阈值对不同编码方法的影响	45
3.3.3	基于提示图和模型的二维初始分割结果	46
3.3.4	基于初始分割和修正点信息融合的二维结果修正	47
3.3.5	与其他二维交互式分割方法相比较	48
3.3.6	分割未见过的二维医学图像	49
3.4	三维医学图像实验验证	51
3.4.1	三维实验数据集	51
3.4.2	基于提示图和模型的三维初始分割结果	52
3.4.3	基于初始分割和修正点信息融合的三维结果修正	53
3.4.4	与其他三维交互式分割方法相比较	54
3.4.5	分割未见过的三维医学图像	54
3.5	讨论与分析	55
3.6	本章总结	58
第四章	基于动态混合伪标签的涂鸦监督医学图像分割	59
4.1	引言	59
4.2	算法设计	60
4.2.1	涂鸦监督学习	60
4.2.2	双分支网络	61
4.2.3	动态混合伪标签	61
4.2.4	目标函数	62
4.3	实验验证	62
4.3.1	实验数据与实现细节	62
4.3.2	算法性能比较	64
4.3.3	消融实验	65
4.4	讨论与分析	69
4.5	本章总结	71

第五章 基于双任务一致性的半监督医学图像分割	72
5.1 引言	72
5.2 算法设计	73
5.2.1 双任务一致性	73
5.2.2 基于双任务一致性的半监督学习	75
5.3 实验验证	76
5.3.1 实验数据及实施细节	76
5.3.2 算法性能比较	77
5.3.3 消融实验	82
5.4 讨论与分析	83
5.5 本章总结	85
第六章 基于主动学习的跨域医学图像分割及临床验证	87
6.1 引言	87
6.2 算法设计	89
6.2.1 基于目标域无关和目标域相关的双参考主动样本选择	89
6.2.2 基于多阶段半监督学习的无源主动域适应	92
6.3 实验验证	93
6.3.1 实施细节	93
6.3.2 多中心鼻咽癌靶区无源主动域适应分割	95
6.3.3 临床可用性分析	102
6.4 讨论与分析	106
6.5 本章总结	108
第七章 总结与展望	109
7.1 全文总结	109
7.2 论文工作创新点	109
7.3 未来工作展望	110
致 谢	113
参考文献	114
攻读博士学位期间取得的成果	134

图目录

图 1-1 不同成像设备针对不同器官或病灶的成像示例	2
图 1-2 图像分割方法主要发展时间线	4
图 1-3 交互式图像分割方法的工作流程图，图片参考自 ^[1]	7
图 1-4 医学图像分割任务中不同图像标注形式的示例	9
图 1-5 主动学习工作流程图，图片参考自 ^[1]	10
图 1-6 论文组织结构	12
图 2-1 图割法中图的建立和图割示意图，图片参考自 ^[2]	16
图 2-2 卷积神经网络结构示意图，以 AlexNet ^[3] 为例	17
图 2-3 卷积操作示意图，以 5×5 图像和 3×3 卷积核和 1 步长为例	18
图 2-4 四种不同激活函数示意图	19
图 2-5 池化操作示意图	20
图 2-6 基于深度学习的图像分割方法发展历程	21
图 2-7 有限标注学习的主要研究问题及方法	22
图 2-8 不同半监督学习策略示意图	27
图 2-9 多模态脑肿瘤分割数据集	29
图 2-10 左心房分割数据集示例	30
图 2-11 不同腹部多器官分割数据集示例	31
图 2-12 腹部器官分割任务中密集标注和稀疏标注（涂鸦）的三维差异示 例，第一行和第二行分别显示了 2D 和 3D 空间中的可视化	32
图 2-13 心脏结构分割任务中像素级标注和稀疏标注（涂鸦）的差异示例。 BG、RV、Myo、LV 和 UA 代表背景、右心室、心肌、左心室和 未标注像素	33
图 2-14 多中心鼻咽癌数据集数据特性图。(a) 不同中心数据的特征分布的 t-SNE ^[4] 可视化图；(b) 不同中心数据的平均灰度分布曲线图。不 同颜色表示数据来自不同医疗中心	33

图 3-1	本章提出的最小交互深度学习分割框架 (MIDeepSeg) 流程图。在第一阶段中 (Stage 1) 用户提供的内部边界点通过指数化测地线距离 (EGD) 进行编码, 以引导模型获得初始分割结果。在第二阶段中 (Stage 2) 根据额外的用户修正点和本章提出的信息融合图割 (IF-GC) 来修正初始分割。请注意, 此框架可用于先前未见目标的分割, 无需额外的微调或重新训练.....	37
图 3-2	训练数据集上不同形状的胎盘进行内部边界点的模拟。紫红色: 在胎盘边缘模拟的点。棕色: 内部边界点导出的边界框。黄色: 金标准	38
图 3-3	从用户提供的内部边界点生成的不同提示图的可视化比较。(紫红色: 内部边界点。棕色: 边界框)	39
图 3-4	修正信息融合示意图。(a) 用户提供点以指示欠分割 (红色) 和过分割 (青色) 区域。(b) 和 (c) 分别是第一阶段 CNN 获得的初始分割前景和背景概率图。(d) 和 (e) 分别是基于前景和背景修正点以及 EGD 转换得到的提示图。(g) 和 (h) 分别是校准的前景和背景概率图。(f) 是修正后的分割结果。其中 IF、EGD 和 GC 分布代表信息融合、指数化测地线距离转和图割	42
图 3-5	二维、三维交互式分割标注界面.....	43
图 3-6	不同数量的初始内边缘点对具有复杂形状的胎盘和前列腺 (未见过目标) 分割的影响。第一行展示了不同数量内边缘点的输入图像, 第二行展示了分割结果	47
图 3-7	不同编码方法在胎盘和脾脏分割的可视化比较, 这是本章方法的第一阶段。第一列显示了带有用户提供的内边缘点 (品红色) 的图像, 其他列显示了初始结果	48
图 3-8	对 GC 和 IF-GC 进行可视化比较。第一行显示了用于修正初始分割结果的用户点。其他行分别显示了通过 GC 和 IF-GC 修正后的结果。这些结果是基于相同的用户点进行修正的。GC 和 IF-GC 分别表示原始图割和信息融合图割	49
图 3-9	MIDeepSeg 对不同目标所需的修正点数量直方图。在第一行中, 胎盘 T2 和脾脏 CT 是见过的目标, 而在第二行中, 脾脏 T2 和肾脏 T2 是以前未见过的目标.....	50

图 3-10 MIDeepSeg 与其他交互式方法在胎盘分割方面的可视化比较。第一行展示了带有或不带有初始交互的初始分割结果。第二行显示了修正后的最终结果.....	51
图 3-11 MIDeepSeg 对二维未见过器官分割结果的一些示例。第一行显示了初始用户交互和初始分割。第二行显示了所有用户交互和最终的分割结果。请注意，该模型仅使用 T2 MRI 中的胎盘进行训练	51
图 3-12 不同交互式分割方法针对未见过的目标的 Dice、ASSD、用户时间和交互数量的比较。其中， $\sharp$ 、 $\S$ 、 $\spadesuit$ 、 $\clubsuit$ 分别表示 CT 图像、超声图像、T1 和 T2 图像。所有这些器官在训练集中都未出现过.....	52
图 3-13 第一阶段不同编码方法分割肿瘤核心结果的可视化比较。所有这些方法都使用相同的内部边缘点和边界框作为输入	53
图 3-14 不同修正方法对三维肿瘤核心分割的可视化比较。这些修正方法在相同的初始分割和相同的修正点条件下进行了比较	55
图 3-15 MIDeepSeg、DeepIGeoS、3D Graph Cuts 和 ITK-SNAP 对肿瘤核心分割结果的可视化比较.....	56
图 3-16 使用 MIDeepSeg 分割未知对象的三个示例。请注意，训练集仅包含 T1C 图像中的肿瘤核心数据	57
图 3-17 不同交互式分割方法对于三维未见过目标分割的 Dice 和 ASSD 及交互时间	57
图 4-1 本章提出方法的概述。该方法由一个编码器 (θ_e)、主解码器 (θ_{d1}) 和辅助解码器 (θ_{d2}) 组成，并分别使用涂鸦注释进行训练 (L_{pCE})。同时，通过动态混合两个解码器的输出生成伪标签，并将其用作进一步网络训练的伪标签监督 (L_{PLS})	60
图 4-2 提出方法与其中现有七种弱监督学习方法在心脏结构分割上的可视化比较.....	68
图 4-3 本章方法与其他弱监督方法在 WORD 数据集上分割结果的可视化比较	68
图 4-4 模型在不同 λ 下的分割结果曲线.....	69

图 5-1	用于半监督医学图像分割的双任务一致性算法流程图。该网络由像素级分类分支（任务 1）和水平集函数回归分支（任务 2）组成，采用了广泛使用的编码器-解码器网络作为分割网络，即 VNet ^[5,6] 。该模型通过最小化有标注数据上的全监督损失函数 L_{Dice} 、 L_{LSF} 和无标注数据以及有标注数据上的双任务一致性损失 L_{DTC} 进行模型优化。 T 函数用于金标准转换成水平集表示以进行监督训练。 T^{-1} 函数将水平集表示转换为分割结果，以计算双任务一致性损失 L_{DTC}	74
图 5-2	不同医学图像半监督分割方法在 20% 标注数据下模型预测结果的 3D 可视化比较。第一行是腹部胰腺分割结果，第二行是左心房分割结果	82
图 5-3	腹部胰腺分割数据集在不同训练方法情况下预测结果的 3D 可视化比较。该实验仅使用 12 个有标签图像进行训练，没有使用未标记的图像。Seg: 单一分割任务的预测结果，LSF: 单一水平集表示回归结果，Seg+LSF: 双任务网络中分割分支的预测结果，Seg+LSF+DTC: 基于双任务一致性的双任务网络中分割分支的预测结果，GT: 金标准.....	83
图 5-4	双任务一致性方法在不同标注数据比例下的腹部胰腺分割性能曲线图。红色虚线和绿色虚线分别展示了全监督 VNet 和双任务 VNet 的分割结果，它们仅使用了标记数据进行训练。蓝色实现展示了基于双任务一致性的半监督分割结果	84
图 6-1	传统多中心迁移学习与本章提出的无源主动域适应（SFADA）学习的流程差异。本方法旨在减少图像标注的同时避免源域数据泄露..	88
图 6-2	本章提取主动样本选择策略 STDR 的流程概述。给定源域图像 I^s 以及标注 Y^s ，STDR 首先在源域数据中训练分割网络。然后将分割特征映射到隐空间，得到 F^s ，聚类算法聚类出参考点 R^s 。目标域中的数据经过类似的操作，得到 F^t 。然后选择两种类型的代表性样本，由放射肿瘤学家手动标记	90
图 6-3	源域（SPH）和目标域（APH）中目标域无关和目标域相关样本可视化比较。源域图像与目标域无关图像相似性更高，而与目标域相关图像的差异更明显。STDR 为本文提出的源域目标域双参考选择策略	91
图 6-4	多阶段半监督学习算法流程	93

图 6-5 不同域适应方法的可视化比较。第一行和第三行展示了二维的可 视化结果，第二行和第四行提供了三维结果.....	101
图 6-6 不同域适应方法在多标注者分割数据集上的可视化比较。最前四 行展示了二维切片的预测轮廓，后四行展示了三维区域预测结果 ...	104
图 6-7 SMU 到 SPH 和 WCH 时，肿瘤不同 T 分期之间的准确性比较.....	107
图 7-1 小病灶分割示意图，第一行为颅脑转移瘤分割任务，第二行为颈 部淋巴结分割任务	112

表目录

表 2-1	本文使用到的多中心腹部多器官分割数据集.....	30
表 2-2	多中心和多标注者鼻咽癌靶区勾画数据集详细临床信息。* 表示 数值结果为中位数和范围	34
表 2-3	多标注者鼻咽癌靶区勾画数据集的相关标注者信息	35
表 2-4	不同肿瘤科医生在勾画鼻咽癌原发灶时的差异，以 Dice(%) 和 95HD (<i>mm</i>) 为例	35
表 3-1	用于训练和测试二维交互式分割算法的数据集。注意，对于脾脏， BTCV 和 TCIA 分别是训练集和测试集	45
表 3-2	使用同一组真实用户交互点，对欧式距离变换的不同阈值进行胎 盘和脾脏分割的定量评估	45
表 3-3	使用同一组真实用户交互点，对高斯变换的不同 σ 值进行胎 盘和脾脏分割的定量评估	46
表 3-4	使用同一组真实用户交互点，对测地线距离变换的不同阈值进行 胎盘和脾脏分割的定量评估	46
表 3-5	使用同一组内边缘点进行胎盘和脾脏分割时不同编码方法的定量 比较。* 表示与第二名方法相比的 p -值 < 0.05	47
表 3-6	使用相同修正点进行胎盘和脾脏分割结果修正的定量比较。* 表 示 IF-GC 与 GC 结果存在显著差异 (p 值 < 0.05)	48
表 3-7	不同交互式方法在 Dice 系数、平均表面距离 (ASSD)、用户时间 和交互点数量方面对二维胎盘和脾脏分割的定量比较。* 表示与 第二名方法相比的 p -值 < 0.05	49
表 3-8	用于训练和测试的三维实验数据集	53
表 3-9	在相同的内部边缘点和边界框情况下，使用不同编码方法进行肿 瘤核心分割的定量比较。结果基于 MIDeepSeg 的初始分割（第 1 阶段）。* 表示与第二名方法相比 p -值 < 0.05	54
表 3-10	不同修正方法在相同的修正点下对三维肿瘤核心分割的定量比 较。修正前的分割是在第一阶段由 MIDeepSeg 获得的。GC: 3D Graph Cuts; IF-GC: 信息融合图割; * 表示性能显著高于 GC, p 值 < 0.05	54

表 3-11 不同交互式分割方法在肿瘤核心分割任务上 Dice 和 ASSD 的定量评估。* 表示与第二名方法比较时的 p 值 <0.05	56
表 4-1 本章提出方法在 ACDC 数据集上与现有弱监督/半监督方法在 DSC 和 95HD (mm) 的结果比较。所有结果均基于相同骨干结构 (2D UNet) 的 5 折交叉验证。* 表示本章方法与第二名方法 (RLoss ^[7]) 配对 t 检验时 p 值 <0.05 。WSL, SSL 和 FSL 分别表示弱监督学习, 半监督学习和全监督学习	66
表 4-2 不同弱监督分割方法在 WORD 数据集上的 Dice(%) 和 95HD (mm) 指标的定量比较。所有结果均基于相同骨干结构 (3D-UNet)。* 表示本章方法与第二名方法 (RLoss ^[7]) 配对 t 检验时 p 值 <0.05	67
表 4-3 双分支网络对不同监督策略的消融实验研究。Single 表示仅使用 pCE 的 UNet ^[8] 。CR 表示不同解码器分支之间的一致性正则化 ^[9] 。CPS 是 Chen 和 Wu 等人 ^[10,11] 提出的交叉伪监督策略。Ours 是本章提出的 PLS 策略。 θ_{d1} 和 θ_{d2} 分别表示主要和辅助解码器分支的预测	70
表 5-1 在腹部胰腺 CT 数据集上, 本章提出方法与其它医学图像半监督分割方法的定量比较。第一行和第二行是本实验的全监督学习结果的下限和上限, 最后一行是本章提出的方法, 其它是以前的方法 ..	79
表 5-2 左心房 MRI 数据集上, 本章方法与其它医学图像半监督分割方法在不同标注比例下的定量比较。第一, 二行和第三行是本实验的全监督学习结果的下限和上限, 最后一行是本章提出的方法, 其它是以前的方法	80
表 5-3 本章方法的不同组成部分在腹部胰腺 CT 数据集上的消融实验	81
表 6-1 U-Net 在不同数据集上鼻咽癌靶区分割性能的定量比较结果	94
表 6-2 SPH 和 SMU 到其他中心的直接预测结果的定量比较	95
表 6-3 使用 SPH 数据集训练并适应到其他数据集上的定量结果。* 表示在所有配对的 t 检验中 p 值 < 0.05 , 表明该方法显著优于其他方法 ..	96
表 6-4 使用 SMU 数据集训练并适应到其他数据集上的定量结果。* 表示在所有配对的 t 检验中 p 值 < 0.05 , 表明本方法显著优于其他方法 ..	97
表 6-5 SPH 数据集到其他数据集样本选择方法的定量结果。* 表示在所有配对的 t 检验中 p 值 < 0.05 , 表明本方法显著优于其他方法	98

表 6-6 SMU 数据集到其他数据集样本选择方法的定量结果。* 表示在所有配对的 t 检验中 p 值 < 0.05 ，表明本方法显著优于其他方法	99
表 6-7 SPH 数据集到其他数据集的消融实验结果.....	100
表 6-8 SMU 数据集到其他数据集的消融实验结果.....	101
表 6-9 不同主动样本数量下，本章方法的多中心泛化结果	102
表 6-10 SPH 数据集训练的模型适应各个标注专家结果的定量分析。* 表示在所有配对的 t 检验中 p 值 < 0.05 ，表明该方法显著优于其他方法	103
表 6-11 SPH 到 APH 实验设置下，STDR 方法中聚类数量对实验结果的影响	105
表 6-12 不同数据集中专家修正程度的定量结果	106

缩略词表

英文缩写	英文全称	中文全称
2D	Two Dimension	二维
3D	Three Dimension	三维
95HD	95% Hausdorff Distance	95%豪斯多夫距离
ACDC	Automated Cardiac Diagnosis Challenge	心脏自动诊断分割比赛
Adam	Adaptive Moment Estimation	自适应矩估计
AL	Active Learning	主动学习
APH	Anhui Provincial Hospital	安徽省省立医院
ASPP	Atrous Spatial Pyramid Pooling	空洞空间金字塔
ASD	Average Surface Distance	平均表面距离
ASSD	Average Symmetric Surface Distance	平均对称表面距离
BP	Back Propagation	反向传播
BraTS	Brain Tumor Image Segmentation	脑肿瘤分割
CAD	Computer Assist Diagnosis	计算机辅助诊断
CAM	Class Activation Mapping	类别激活图
CE	Cross-Entropy	交叉熵
CRF	Conditional Random Field	条件随机场
CT	Computed Tomography	计算机断层扫描
CNN	Convolutional Neural Network	卷积神经网络
DBN	Dual-Branch Network	双分支网络
DMPLS	Dynamically Mixed Pseudo Labels Supervision	动态混合伪标签监督

DTC	Dual-Task Consistency	双任务一致性
EMA	Exponential Moving Average	指数移动平均
Eucl	Euclidean Distance Transformation	欧式距离变换
EGD	Exponentialized Geodesic Distance	指数化测地线距离变换
FCN	Fully Convolutional Network	全卷积网络
FC	Fully Connection	全连接层
IF-GC	Information Fusion followed by Graph Cuts	信息融合图割
GC	Graph Cuts	图割
Gauss	Gaussian Transformation	高斯变换
GD	Gradient Descent	梯度下降
Geos	Geodesic Distance Transformation	测地线距离变换
GMM	Gaussian Mixture Model	高斯混合模型
GTV	Gross Tumor Volume	原发灶
HITL	Human-In-The-Loop	人在回路
HU	Hounsfield Unit	亨氏单位
LA	Left Atrium	左心房
LR	Learning Rate	学习率
LSF	Level Set Function	水平集函数
LV	Left Ventricle	左心室
MIDeepSeg	Minimally Interactive Segmentation using Deep Learning	最小化交互的深度学习交互式分割算法
MHSA	Multi-Head Self Attention	多头注意力
MRI	Magnetic Resonance Imaging	核磁共振成像
Myo	Myocardium	心肌

MSE	Mean-Square Error	均方误差
NSD	Normalized Surface Distance	归一化表面距离
NPC	Nasopharyngeal Carcinoma	鼻咽癌
OCT	Optical Coherence Tomography	光学相干断层成像
PET	Positron Emission Tomography	正电子发射断层成像
PLS	Partially Label Supervision	伪标签监督
ReLu	Rectified Linear Unit	修正线性单元
RF	Random Forest	随机森林
RV	Right Ventricle	右心室
SCH	Sichuan Cancer Hospital	四川省肿瘤医院
SFDA	Source-Free Domain Adaptation	无源域适应
SFADA	Source-Free Active Domain Adaptation	无源主动域适应
SGD	Stochastic Gradient Descent	随机梯度下降
SMU	Southern Medical University	南方医科大学
SPH	Sichuan Provincial People Hospital	四川省人民医院
SSL	Semi-Supervised Learning	半监督学习
SVM	Support Vector Machine	支持向量机
STDR	Source-Target-domain Dual-Reference	源域目标域双参考
USI	Ultrasound Imaging	超声成像
WSL	Weakly-Supervised Learning	弱监督学习
WCH	West China Hospital	华西医院

第一章 绪论

1.1 课题背景及研究意义

医学图像分割通常是指从医学图像数据中将相关结构、组织或病变提取出来的过程。随着成像技术的快速发展及其临床应用，尤其是超声成像（Ultrasound Imaging, USI），X射线（X-Ray），计算机断层成像（Computed Tomography, CT），核磁共振成像（Magnetic Resonance Imaging, MRI），正电子发射断层成像（Positron Emission Tomography, PET），光学相干断层成像（Optical Coherence Tomography, OCT）等先进成像技术的不断进步和普及，医学图像数据在数量和多样性方面也在快速增加^[12-14]。比如，CT以在高分辨率、多平面重建、快速成像、对软组织和骨组织的优秀成像能力被广泛用于骨折等疾病诊断，损伤评估和介入治疗。MRI以优秀的软组织对比度、从不同角度揭示组织的病理和生理状态、观察大脑功能活动神经网络连接和认知过程、以及可造影实现局部增强等成像特点被广泛应用于临床诊断，疾病监测和认知研究。PET以非侵入式、高灵敏度、可定量分析、可融合成像和功能成像等特点被广泛引用于肿瘤监测，生物学参数分析和癌症标志物监测等。不仅如此，由于疾病特性、成像特点和临床诊断治疗需求的不同，实际临床产生的医学图像也存在着巨大差异，如图1-1所示。随着这些多样化临床数据的快速增长，计算机辅助诊断（Computer Assist Diagnosis, CAD）被提出以加速医学图像处理过程^[15,16]。作为计算机辅助诊断系统不可替代的组成部分，医学图像分割被广泛用于帮助医生从图像中准确地分割出感兴趣的结构或区域辅助诊断、治疗和预后分析等，加速临床决策流程的同时降低医生的工作负担。

到目前为止，医学图像分割受到了广泛的关注，并取得了长足进步。早期阶段，由于计算机技术和成像技术尚未发展成熟，医学图像的获取和处理都相对比较困难，因此分割过程主要依靠医生的经验和直觉进行手动完成，缺乏系统化的算法支持。随着计算机技术的发展，基于阈值的分割方法开始得到应用^[17]。该方法基于像素灰度分布直方图来选择阈值将图像分割成不同的区域。这种方法简单易用，但对于复杂和灰度分布不均匀的医学图像效果不佳。之后，区域生长、区域分裂合并和区域分割等基于区域的分割方法逐步被提出和应用用于医学图像分割任务^[18-20]。这些方法通过将相邻像素组合成区域，并根据一些特定的规则将图像分割为不同的区域。此外，边缘检测方法也被用于识别医学图像中的边界和轮廓^[21]。尽管这些方法取得了一定进步，由于分割性能不佳依旧很难用于实际临床中。二十一世纪以来，随着统计和机器学习的快速发展，基于图像特征分割方法变得流

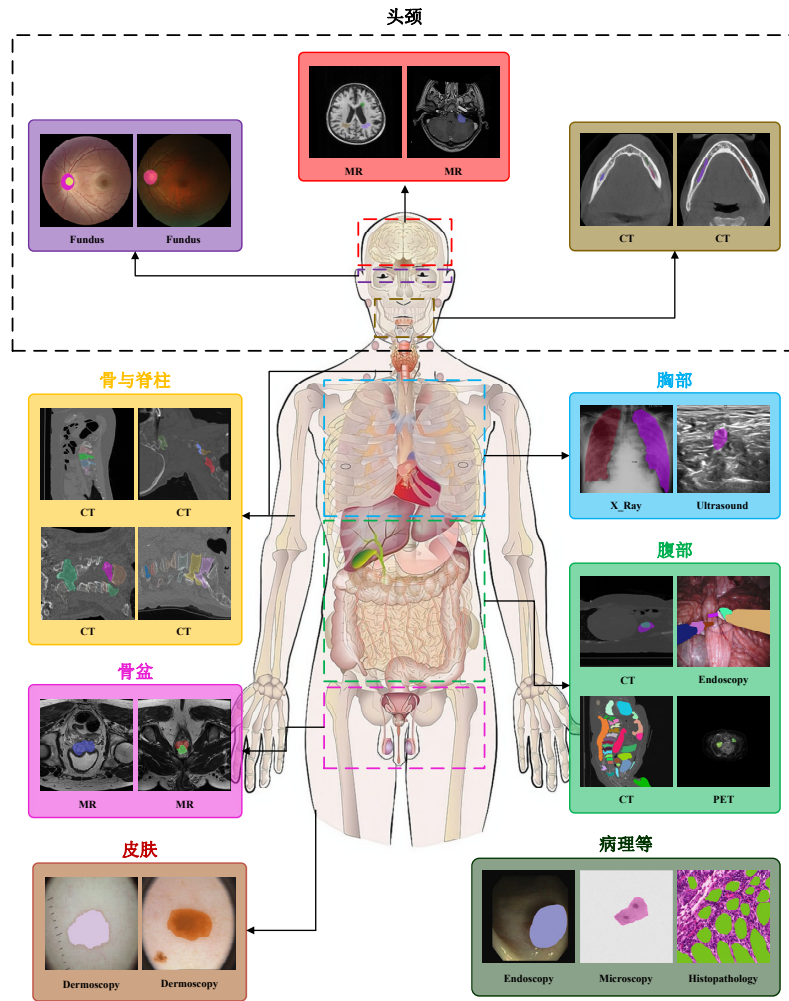


图 1-1 不同成像设备针对不同器官或病灶的成像示例

行起来。这些方法利用机器学习算法，如图割^[2]（Graph Cuts, GC）、条件随机场^[22]（Conditional Random Field, CRF）、支持向量机^[23]（Support Vector Machine, SVM）、高斯混合模型^[24]（Gaussian Mixture Model, GMM）、随机森林^[25]（Random Forest, RF）等，从图像中学习特征并进行分割。此外，基于曲线演化的活动轮廓模型^[26]（Active Contour Model, ACM）、基于偏微分方程的水平集^[27]（Level Set）、基于互信息^[28]（Mutual Information）的配准也被广泛用于图像分割。这些方法多为半自动分割方法，需要用户为算法提供初始种子点用于图像分割。

最近，随着深度学习的快速发展，卷积神经网络（Convolutional Neural Network, CNN）被广泛用于从大量标注数据中自动进行特征学习来理解图像^[3,29]，并获得了巨大成功。例如，基于残差神经网络的深度学习模型在图像分类任务的 ImageNet 数据集上准确率首次超越了人类^[30]。全卷积网络^[31]（Fully Convolutional Network, FCN）大幅刷新了多个语义分割任务的结果。在医学图像分析方面，基于全卷积

网络的 UNet 网络^[8] 及其变体结构^[32-34] 在多个分类、分割检测任务上取得了可喜的进展，部分任务上的精度甚至可与初级医师媲美^[35,36]。这些结果说明基于数据驱动的深度学习方法可以从大量标注数据中自动学习并生成与标注相似的预测并可进一步被实际应用，为加速计算机辅助诊断提供了新的可行解决方案。尽管如此，最新研究发现受到数据规模和多样性差异等影响，基于数据驱动的深度学习方法在实际应用场景中也遇到了许多挑战^[16,37-42]，主要表现如下：

1) 数据标注成本高：在深度学习模型开发过程中，收集和标注大量数据是一个成本极高且耗时的过程，尤其是图像分割任务。不同于分类任务只需几秒钟即可给出图像的类别信息，分割标注需要为图片中每个像素提供类别信息，根据图像大小和类别多少的不同，获取一张高质量标注通常需要耗费数小时。例如，Cordts 等人^[43] 的研究表明标注一张二维街景图像平均需要花费 1.5 小时。不同于二维自然图像，医学图像的标注成本更高且更具挑战性。首先是医学图像标注需要专业的领域知识和临床经验，因此标注过程只能由经验丰富的专科医生进行。此外，在医学影像标注时医生往往需要参考前后层器官解剖结构变化趋势，且由于对比度较差，医学图像标注过程较自然图像更加耗时，Luo 等人^[44] 的研究表明，标注一个全腹部三维 CT 影像的 16 个器官平均需要四小时。

2) 数据分布差异大：此外，不同医院的影像数据普遍存在分布差异，对于图像分割来说这些差异主要存在两个方面。第一是成像差异，受不同成像协议、参数、设备供应商等影响，不同医院的影像数据存在较大的分布差异。第二是标注差异，受到教育经历、临床经验、临床指南理解等差异，不同医生对器官或病灶的标注也存在着较大差异^[45]。同时，由于医学数据的敏感性，隐私保护使得不同医院数据和标注共享变得困难。这些差异为深度学习模型的实际应用带来了新的挑战，即如何利用有限标注数据来提高深度学习模型在数据分布差异较大任务上的准确性和泛化性^[39,46]。在实际应用过程中，由于医学图像数据的多样性，针对特定成像或病种开发相应自动分割模型时均需要标注大量数据，显著增加了模型的开发成本。

针对上述主要挑战，本研究旨在**探索如何利用有限标注数据来训练高性能、高泛化性的图像分割模型，降低数据标注成本的同时减轻医生的标注负担**，这也是现阶段医学图像分析领域关注的重点。缓解医学图像分割中深度学习模型对标注数据的依赖，减低模型开发和应用的标注成本是计算机辅助诊断系统面临的实际问题，可以提高深度学习模型实际临床应用的前景，并进一步为临床诊断、治疗和研究提供有价值的信息用于决策。尤其是缓解模型对标注数据的依赖可以提高计算机辅助诊断系统开发的效率，减少模型开发和部署的等待时间。与此同时，

利用有限标注数据也可以更加高效的实现跨医院和跨医生的知识迁移，进而提高计算机辅助诊断系统在实际场景的鲁棒性，为实现跨医院同质化诊疗提供可能。通过有限标注数据和深度学习优越性能，使得为各种医学图像低成本快速开发自动分割模型成为可能，促进医疗资源共享的同时推动智慧诊疗的快速发展。

1.2 国内外研究现状

1.2.1 图像分割发展现状

图像分割方法有着悠久的历史，从方法学角度来说，可以简要将其分为基于规则、基于模型、基于优化和基于学习四个类别。图1-2简要的列出了这些方法的发展历程。

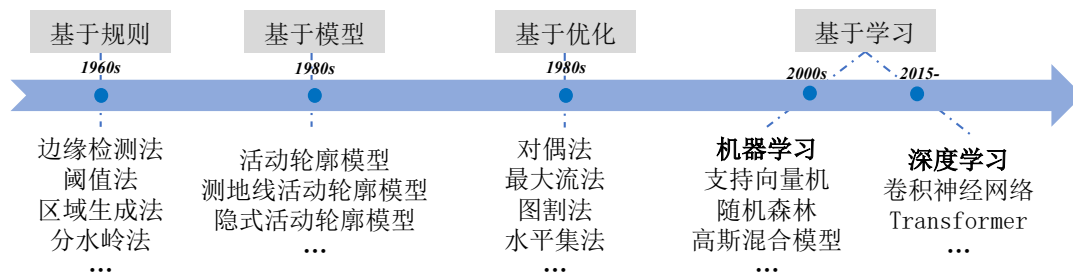


图 1-2 图像分割方法主要发展时间线

1) 基于规则的图像分割方法: 上世纪六十年代到八十年代，图像分割主要为基于规则的方法。比如，边缘检测法，通过检测图像中的边缘或轮廓实现物体的分割，主要方法包括 Sobel^[47]、Prewitt^[48]、Canny^[49] 等算子以及边缘检测滤波器。此外，阈值法^[17,50,51] 根据图像中的像素灰度或颜色分布，设置阈值将像素分为不同的类别，高于阈值的区域则被视为目标前景。区域生长法^[20] 根据给定种子点的灰度信息将图像分为具有相似特征或属性的区域（前景）和不相似的区域（背景）。分水岭法^[52] 将图像视为地形地貌，图像中的每个像素点被看作是地形上的一个点，该点的灰度值被看作为海拔高度。分水岭算法根据海拔高度信息将图像分隔为不同区域并形成不同的“山峰”和“山谷”，然后将每个“山谷”或“山峰”内的像素视为同一区域，这些不同区域即为分割结果。综上，基于规则的图像分割方法实现简单，计算复杂度低，但只能处理对比度清晰的简单图像。对于背景复杂，图像噪声较大的医学图像，此类方法很难取得较好的分割结果，难以实际应用。

2) 基于模型的图像分割方法: 上世纪八十年代，随着计算机和应用数学的快速发展，基于能量泛函和曲线演化的活动轮廓模型被提出和广泛研究^[53]。活动轮廓模型的基本思想是定义一个能量函数来描述分割物体的轮廓曲线在图像中的形变过程，并使其能够自动地沿着初始化曲线收缩或扩张，最终收敛到分割目标的边缘

处。这个能量函数由内部能量项和外部能量项组成，外部能量项惩罚曲线的曲率和长度，内部能量项则根据图像的灰度分布或其他特征引导曲线向感兴趣对象的边界移动。在此基础上，根据分割目标的几何特点引入正则化以加速轮廓演化和提高分割效果也获得了广泛的关注^[26,54,55]。基于模型的分割方法可解释性极强且能够根据分割目标的几何特点来设计特定的约束项保证其解剖结构的稳定性，被广泛的应用到医学图像分割任务中。如著名的开源医学图像处理软件 ITK-SNAP^[56]和其他分割工具^[57]均采用了此方法来进行图像分割。但该类方法对模型参数、初始输入以及迭代过程非常敏感，需要针对每个图像单独设计，很难普遍适用。

3) 基于优化的图像分割方法: 基于能量泛函分割方法的快速发展，激发了图像分割领域的持续进步，其中最重要的影响就是基于优化方法的图像方法获得了普遍关注。基于优化的图像分割方法通过定义一个或者一组优化目标函数，并通过优化算法来最小化或最大化目标函数以达到图像分割的目的。这些方法通常将图像分割问题转化为一个优化问题，通过调整分割结果的变量（如分割边界或区域类别）来优化目标函数的值，如原对偶法^[58]、最大流最小割法^[59]、图割法^[2]、水平集法^[60]。基于优化的图像分割方法基于应用数学有很强的理论保证^[61]，并在多个应用场景下取得了良好的表现^[2,62,63]。尽管如此，这些方法也受限于参数多和泛化性差等缺点，难以直接批量处理医学影像。

4) 基于学习的图像分割: 随着机器学习和深度学习的快速发展，基于学习的方法逐步发展成为图像分割方法的主流。在深度学习快速流行之前，机器学习通过将图像中的每个像素视为一个样本，通过训练一个分类器来判断每个像素属于哪个类别来实现图像分割，如支持向量机^[23]、随机森林^[64]、高斯混合模型^[65]等。这类方法能直接从标记好的数据中学习图像的特征和分割规律，然后将学习到的知识应用到新的图像中进行分割，大幅提高了图像的准确性和效率。但这些传统机器学习方法通常需要手动设计特征和参数，且在处理复杂图像和场景时表现不佳。随着卷积神经网络 (Convolutional Neural Networks, CNN) 在图像分类任务上获得突破性成功^[3]，基于 CNN 的深度学习获得了持续关注并被快速应用于图像分割任务。通常分类任务的 CNN 由多个卷积 (convolution)、池化 (pooling)、全连接层 (Fully Connected Layers, FC) 组成，通过为每个图像输出一个向量来代表预测结果。不同于分类任务，分割任务需要为每个像素输出一个分类结果，因此最后的预测结果为与原图大小一致的分类结果。为了实现这一目的，Long 等人^[31]设计了全卷积网络用于图像分割，该网络采用了一个包含编码 (Encoder) 和解码 (Decoder) 的两阶段网络架构并移除了分类任务网络中的全连接层来处理任意尺寸的输入图像。编码阶段，FCN 利用多个卷积和池化操作提取图像在不同尺度下的特征信息，

在解码阶段 FCN 利用线性插值或反卷积将多尺度特征和预测进行逐步上采样直到获得与原图分辨率大小一样的预测结果。受到 FCN 的启发, Ronneberger 等人^[8]将跳跃连接与基于编码和解码结构的 FCN 相结合提出了用于医学图像分割的 U-Net 模型, 并在当年的细胞结构分割比赛中获得了第一名。随后 Cicek 和 Milletari 等人分别通过引入三维卷积将 U-Net 进一步扩展到三维分割任务并开发了广泛应用于三维医学图像分割的 3D U-Net^[32] 和 V-Net^[5]。从此医学图像分割进入了卷积神经网络的时代, 并在一众医学图像分割任务中取得了最好的性能^[16, 34, 66]。最近, Transformer 以其多头自注意模块 (Multi-Head Self Attention, MHSA) 不受感受野约束的优越全局建模能力被广泛用于图像分割^[67]。Zhou 和 Tang 等人分别用多头自注意模块代替 U-Net 中全卷积模块开发了用于医学图像分割的 SwinUNETR^[68] 和 nnFormer^[69], 并在多个医学图像分割任务上取得了最好的分割结果。

尽管深度学习模型在多个医学图像分割公开数据集上取得了优越的性能, 但其性能严重依赖标注数据的数量、质量和多样性^[1, 16, 70]。然而收集和标注大规模高质量的医学图像分割数据成本昂贵也不太现实。Tajbakhsh 等人^[70]系统性地回顾了 100 多项已有工作后指出利用有限标注数据训练模型可以降低深度学习模型对标注数据的依赖。下面将介绍基于有限标注数据的医学图像分割算法的有关研究进展。

1.2.2 高效数据标注方法研究现状

数据标注是开发高性能深度学习分割模型的基础, 受限于医学图像标注成本极高, 如何高效的进行数据标注逐步成为了研究热点^[1]。针对这个问题, Samuel 等人^[1]的最新研究表明基于自动分割和迭代修正的图像标注策略比纯手工勾画需要更少的时间和交互量, 并证明了交互式分割是一种更高效的数据标注方法。交互式分割通过将算法和用户相结合利用用户交互来迭代地修正模型预测, 提高数据标注效率的同时降低数据的标注成本, 该过程也被称为人在回路 (Human-in-the-loop, HITL)。如图1-3所示, 交互式分割工作流程首先需要少量标注数据训练初始图像分割模型, 然后利用该分割模型为未见过的图像产生初始预测, 并将该结果提供给用户进行修正, 然后模型根据用户的修正信息再次更新预测结果, 用户可以持续重复以上步骤直至认为修正结果可接受, 即可保存分割结果用于后续模型迭代。

在图像分割的发展进程中, 许多传统方法都是交互式分割方法, 例如 Graph Cuts^[2]、ITK-SNAP^[56]、GeoS^[71]、Random Walks^[72] 和 GrowCut^[73]。这些方法可用于对医学图像的标注, 但大都使用低级特征 (例如灰度级或颜色分布) 进行图像分割^[74]。由于低级特征在许多低对比度情况下不能有效地区分目标和背景^[74], 这

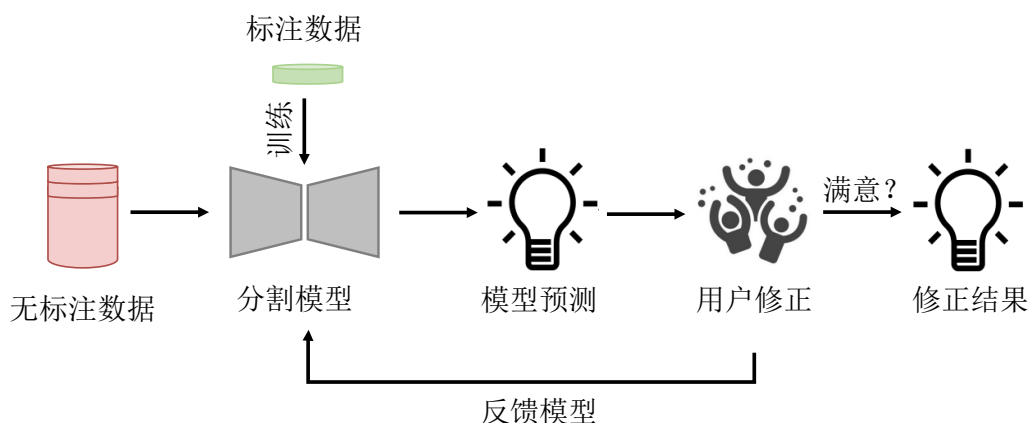


图 1-3 交互式图像分割方法的工作流程图，图片参考自 [1]

些方法通常需要大量的用户交互和长时间才能获得可靠的结果。为了减少用户获取准确前景/背景所需的人工交互量，机器学习被广泛用于交互式分割任务中。例如，SlicSeg^[64] 和 DyBaORF^[25] 使用在线随机森林来从胎儿磁共振图像中分割胎盘。GrabCut^[24] 使用高斯混合模型来估计前景和背景的分布。它通过用户提供的感兴趣区域周围的边界框获得初始结果，并允许进行额外的交互以进行修正。此外 Wang 等人^[75] 还利用主动学习来主动选择候选区域询问用户以获得更多的反馈信息，从而减少用户交互。这些算法比传统方法表现更好，但它们仍受到手工设计特征的限制^[76,77]，仍然需要大量的用户交互来进行准确的分割。最近，随着深度学习及卷积神经网络（CNN）的快速发展，CNN 已经在图像分割方面取得了最先进的性能^[78,79]。为了充分利用 CNN 的良好表征能力，并克服自动 CNN 的准确性和鲁棒性有限等问题，最近提出了一些基于深度学习的交互式分割工具^[74,76,77,80-83]。这些方法^[74,81,82] 旨在对二维自然图像进行交互式分割，缺乏对低对比度和边界模糊的医学图像的评估。此外，一些方法将强化学习和图神经网络集成到统一的基于多边形的交互式分割框架中，允许用户拖动多边形上的点进行修正，但其处理复杂形状的三维医学图像能力有限^[84-86]。DeepIGeoS^[77]、IFSeg^[83]、DeepCut^[80] 和 BIFSeg^[76] 专门设计用于医学图像分割的交互式分割方法。尽管这些方法易于使用，但依旧缺乏处理不同模态中未见过目标的能力。BIFSeg^[76] 利用用户提供的边界框和图像特定的微调来分割一些简单的未见目标，但它依旧不能处理跨模态的数据并且需要在测试时对每个测试图像进行微调。

1.2.3 基于有限标注的学习方法研究现状

最近，Tajbakhsh 等人^[70] 系统地回顾了有限标注学习话题目前存在的主要问题和相关解决办法，并指出基于有限标注数据的医学图像分割是一个重要、热门

且具有挑战性的话题。受此综述的启发,针对标注据有限情况下存在的三个主要问题(如何利用稀疏标注、如何利用未标注数据和如何选择标注数据)开展研究。本小节将介绍每个问题及其主要解决方法的最新进展。从方法学的角度来说,上述三个问题分别对应了三种机器学习方法:弱监督学习(Weakly-Supervised Learning, WSL)、半监督学习(Semi-Supervised Learning, SSL)、主动学习(Active Learning, AL)。下面针对每个方法分别介绍其代表性工作。

1) 弱监督学习:全监督学习需要像素级的强监督信号,但图像标注耗时又乏味,尤其是对于三维医学图像。为了减少数据标注成本,最近一些工作探索利用有限的弱标注数据以弱监督学习模式来训练深度学习模型^[70]。图1-4展示了不同弱标注类型与全标注的标注量差异。由图可知,弱标注类型可以被归纳为两类:1)基于图像类别信息的标注;2)仅标注每张图像中的极少部分或像素的稀疏标注,如,点、涂鸦和框。尽管这些弱标注数据比起全标注数据提供的监督信号有限,但是最近的研究表明使用有限的弱标注数据来训练模型也可以取得不错的分割结果。例如,Lin 等人^[87]提出了一种基于图的标签传播算法方法,把图像中的涂鸦标注传播到未注释的像素获得像素级伪标签,并结合涂鸦标注和伪标签进行模型训练。之后,Tang 等人^[7]引入了条件随机场正则化损失,直接训练分割网络。对于医学图像,Can 等人^[88]提出了一种多阶段迭代框架来利用涂鸦来训练模型。首先,他们将涂鸦作为种子点送入 Random Walks^[72]产生初始分割。然后,使用初始分割结果来训练模型,再使用训练好的模型对图像进行预测并利用条件随机场修正结果进一步产生新的分割结果。最后,多次重复上述步骤来提高模型的分割性能。Kim 等人^[89]提出了基于活动轮廓模型的正则化函数来使用弱标注训练分割网络。Lee 等人^[90]结合了伪标签和标签过滤策略,使用涂鸦标注为模型训练提供可靠的伪标签。Liu 等人^[91]基于不确定性感知的教师-学生模型和变换一致性提出了一个统一的弱监督框架来鼓励模型从涂鸦标注中学习。最近,Valvano 等人^[92]提出了多尺度对抗注意力模型来结合大量涂鸦标注和少量全标注来训练模型。尽管这些方法的分割结果仍然低于全监督模型,但弱标注只需要极少的稀疏标注就可以取得不错的分割结果,很好的展现了降低标注成本的潜力。

2) 半监督学习:半监督学习旨在通过利用少量标注数据和大量未标注数据来开发高性能预测模型,关键步骤是为未标注数据设计有效的监督。目前,许多方法被相继提出用于为未标记数据提供监督信号,这些方法主要可以分为两种:(1)伪标签^[93],该策略首先在标注数据上训练模型并通过预测未标注数据来生成伪标签,然后对这些伪标签进行修正或筛选,最后使用这些修正过的高质量伪标签重新训练模型,并多次重复此策略以更新这些伪标签和分割模型。(2)一致性约束^[94],其

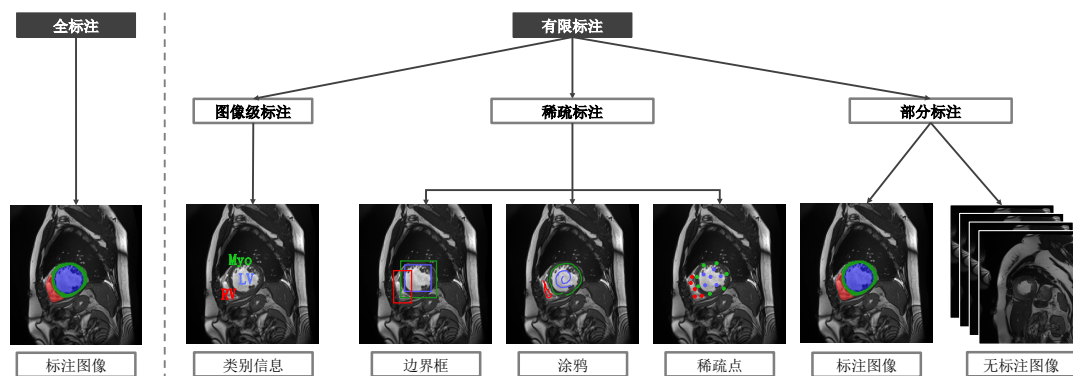


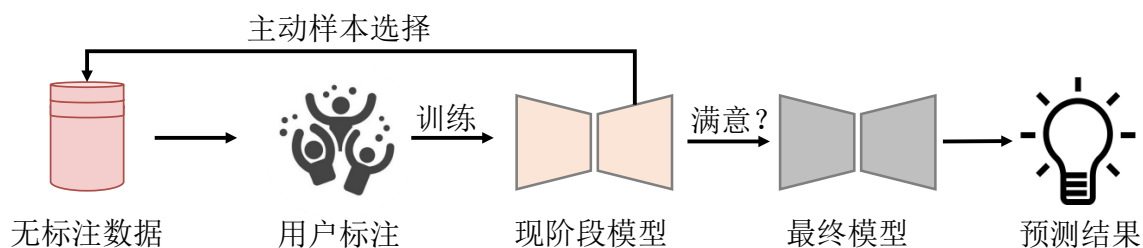
图 1-4 医学图像分割任务中不同图像标注形式的示例

主要假设是模型对于不同扰动下的未标注数据预测结果应该相似，基于一致性约束的半监督学习核心思想是有标注数据用全监督损失、无标注数据用一致性约束联合训练模型。受到这些策略在机器学习上成功应用的启发，许多工作逐步将它们扩展到半监督医学图像分析任务中，包括分类和分割。Bai 等人^[95]提出了一种基于伪标签的迭代学习方法来训练模型从 MRI 图像中分割心脏结构，其核心思想是用模型为无标注数据生成伪标签并用条件随机场来修正伪标签用于下一步训练。Mean-teacher 模型^[94]及其后续改进工作^[96-99]也在半监督医学图像分割中受到越来越多的关注，这些方法通过最小化教师模型与学生模型之间的预测差异来从无标注数据上进行学习。最近工作使用一些新的深度学习技术也在半监督医学图像分割任务上取得了不错的结果，例如深度对抗网络^[100]和注意力机制^[101]。

3) 主动学习: 主动学习是一种高效的平衡数据标注成本和模型分割性能的机器学习方法。如图1-5所示，它通过选择模型认为最有价值的样本进行标注而不是随机选择样本进行标注来降低标注成本^[102]。目前有很多主动学习方法被相继提出用于样本选择，这些策略包括：1) 基于不确定性的方法，重点选择具有高不确定性的样本^[103,104]；2) 基于代表性的方法，优先考虑不同数据分布区域内的样本^[105-108]。这些方法为选择信息量高的样本提供了不同策略，并被广泛用于计算机视觉任务中，如图像分类和图像分割。主动学习具有最小化标注成本的优势，在数据分布相同和标注成本受限的许多医学场景中都取得了优异的表现^[102]。能否利用主动学习来从存在数据分布差异的样本中选择有限的图像进行标注以提高模型的准确性和泛化性是一个重要且未被研究的问题。

1.3 主要研究内容

本论文旨在利用有限标注数据来解决医学图像分割任务中模型对大量标注数据依赖问题的同时降低分割模型开发的标注成本。根据这一研究目的和国内外研

图 1-5 主动学习工作流程图，图片参考自^[1]

究现状，本研究从如何进行高效数据标注、如何利用稀疏标注数据、如何利用少量标注数据和大量未标注数据、如何选择标注数据这四个主要问题出发开展研究。围绕这些问题，本文针对其涉及的主要技术手段中面临的一些困难和挑战，提出了一系列新的、有效的解决方案。本论文的主要研究内容如下：

（1）基于测地线距离变换的可泛化交互式分割标注

本文提出了一种新的基于深度学习和指数化测地线距离变换的交互式医学图像分割方法。该方法以内部边界点作为用户交互，利用指数化测地线距离变换来编码这些交互点为指导图并利用这些交互点裁剪掉图像中多余的背景区域。通过融合指数化测地线距离图来指导深度学习模型为裁剪后的图像生成预测结果。考虑到模型预测结果存在错误需要修正，该方法进一步提出了基于信息融合的图割模型来处理用户提供的修正交互。具体来说该方法再次使用指数化测地线距离变换来编码用户提供的修正信息，并将编码后的图像与模型的初步概率预测相融合，利用条件随机场来更新结果。此外，考虑到单次修正结果难以达到满意，该方法允许用户多次迭代修正，直至修正结果被用户接受。在多中心、多模态和多病种等十余个公开数据集上验证了该方法的有效性。实验结果表明，该方法只需要有限的交互点即可大幅提升分割结果，此外该方法不需要重新训练或微调模型即可直接处理未见过的图像，比已有的交互式分割方法有更好的分割准确度和泛化性。

（2）基于动态混合伪标签的涂鸦监督医学图像分割

本文提出了一个基于双分支网络和动态混合伪标签策略的弱监督学习方法，通过随机混合模型两个分支的输出为无标注像素生成高质量的伪标签，并利用伪标签和涂鸦标注以端到端的形式训练双分支模型。在两个公开医学图像分割基准数据集上的实验表明，本方法比现有的七种弱监督方法取得了更准确的分割结果，证明了该方法的有效性和创新性。

（3）基于双任务一致性的半监督医学图像分割

本文提出了一种新的基于双任务一致性的半监督学习方法，该方法从图像分割表示方法的角度出发，将图像分割定义为两种不同表示（任务），像素层面的分

类任务和全局层面的水平集回归任务。基于这两种表示，本方法用双任务分割网络同时进行分类和回归任务。然后，本方法利用可微的任务转换层讲两个任务映射到同一表示空间，并构建两个任务的一致性来相互约束，约束模型同时关注局部和全局特征。由于双任务一致性仅基于模型预测不需要额外的监督信号，可以直接为无标注数据提供监督信号。在两个公开医学图像半监督分割基准数据集上验证了本方法的有效性，实验结果表明本方法比既往半监督方法取得了更好的分割结果。

(4) 基于主动学习的跨域医学图像分割及临床验证

本文结合主动学习和半监督学习提出了一个新的基于有限标注数据的跨域医学图像分割策略。为了解决由于数据分布导致的模型跨中心性能下降的问题，本文创新性的提出利用主动学习和现有模型在目标域数据中主动选择代表性样本进行数据标注。为了平衡不同中心数据分布差异，本文提出了一个新的双参考样本选择策略，即代表性样本由域无关和域相关样本组成，同时考虑不同中心的数据分布差异提高标注数据的多样性。然后，为了进一步提高模型在目标域的泛化性，本方法进一步采用了多阶段半监督学习利用主动标注样本、无标注数据和半监督学习进行模型微调。该方法只需要源域训练好的模型和目标域中有限标注数据即可实现高性能、高泛化性的图像分割。在来自五个医院的 1057 名患者的多中心鼻咽癌靶区勾画数据集和来自 170 名患者的多标注者数据集验证了该方法的有效性，实验结果表明，该方法显著优化现有域适应和主动学习方法，另外实验结果还表明只用 20% 标注数据即实现与全标注模型相媲美的分割性能。此外，多中心临床验证也表明，在极少的医生介入下该方法实现了临床可用的分割结果。

1.4 本文组织结构

本文围绕如何利用有限标注数据开发高性能医学图像分割模型这一实际问题展开研究，通过现状分析、理论学习，然后针对主要问题提出了四种解决方案，最后总结展望，其组织结构如图1-6所示。全文共分为七章，具体安排如下：

第一章: 绪论，主要介绍了本课题的研究背景和意义，首先简述了医学图像特点和计算机辅助诊断系统发展引出医学图像分割任务的重要性。然后针对医学图像分割所涉及主要技术领域的发展现状进行了回顾总结，分析既往工作存在的问题并引出本文研究方向，即利用有限标注数据来训练高性能深度学习模型，降低模型对标注数据的依赖，进而降低模型开发成本，为本研究提供了理论支撑和应用指导。然后，针对本文研究内容，对有限标注数据相关方法的研究背景和进展进行回顾分析，指出相关技术的优势和不足。最后根据这些优势和不足，提出本

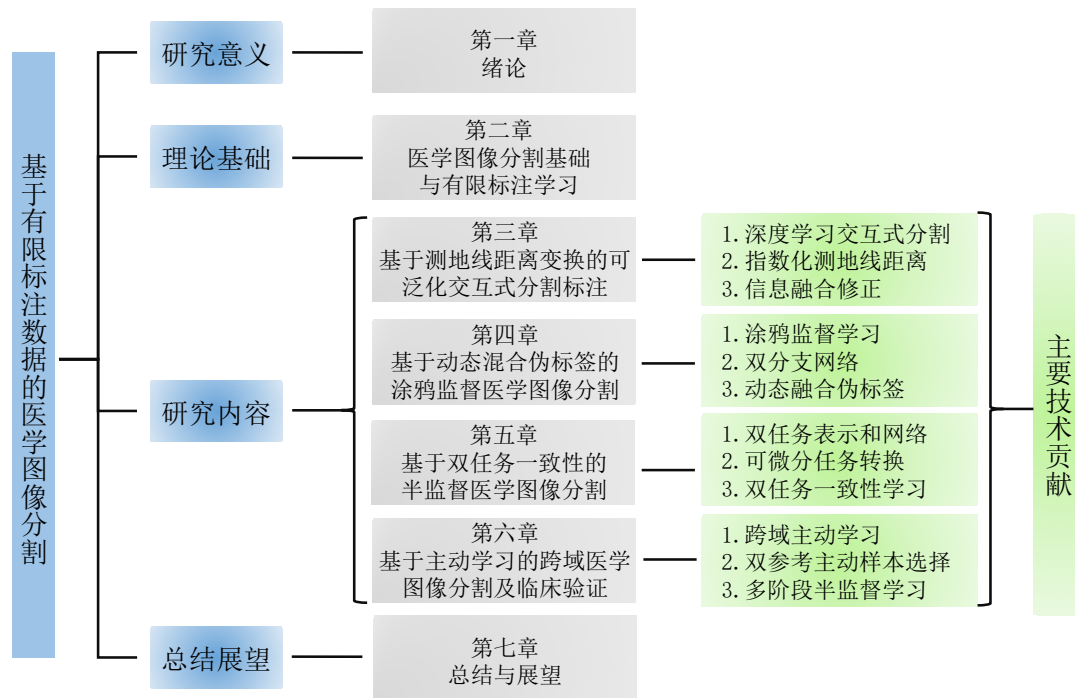


图 1-6 论文组织结构

文的研究路线和相应的解决方案，最后简要总结了本文的组织架构。

第二章：医学图像分割基础与有限标注学习。本章首先介绍了和本文高度相关的传统分割方法。然后详细介绍了深度学习理论基础，包括网络结构、训练和优化策略等。之后介绍了基于深度学习的图像分割方法的发展历程。随后从降低数据标注成本的角度引出了本文的研究重点，有限标注学习。并回顾了既往经典方法的核心技术及其优势和不足。最后介绍了有限标注学习常用的公开基准数据集。

第三章，基于测地线距离变换的可泛化交互式分割标注。首先介绍了交互式医学图像分割方法的相关背景并分析现有方法存在的问题。然后介绍了该方法从交互形式、编码方式和修正方式来提高交互式分割方法准确性、泛化性和易用性，即提出了更易使用和信息最大的内部边界点来作为用户输入；并改进了测地线距离来进行交互编码，提出了信息融合图割法进行结果修正。最后介绍了本方法在多个数据集上的性能表现。

第四章，基于动态混合伪标签的涂鸦监督医学图像分割。首先介绍了涂鸦标注的相关背景和意义及目前常用方法存在的挑战和不足。根据现有问题，本方法提出了双分支网络，通过不同分支产生多样化的预测结果从涂鸦标注上进行学习；然后进一步提出了动态混合伪标签策略，通过随机混合两个分支输出的多样性预测来生成高质量伪标签，最后通过结合高质量伪标签和涂鸦标注以端到端的形式进行模型训练。最后在公开基准数据集上进行实验验证，说明了本方法的有效性。

第五章，基于双任务一致性的半监督医学图像分割。首先对半监督学习的特点和临床意义进行了介绍，紧接着介绍了多任务方法的研究进展和优点。受这些进展和优点的启发，提出了一种双任务一致性方法用于半监督医学图像分割。首先从任务表示形式的角度出发用像素层面的分类和全局层面水平集回归来分别表示图像分割，并进一步提出用双任务网络来分别预测两种表示。考虑到两种表示的互补性，进一步引入了可微分的任务转化函数，将水平集表示转换为像素分割结果，并引入双任务一致性，鼓励两种不同表示方法在分割层面上结果一致。最后通过双任务一致性，模型可以同时利用标注数据和无标注数据进行半监督学习。多个公开医学图像分割半监督数据集上的实验结果表明本方法优于已有半监督医学图像分割方法。此外，该工作也是最早将跨任务一致性引入到半监督学习领域的工作之一。

第六章，基于主动学习的跨域医学图像分割及临床验证。首先对现有基于有限标注数据方法的适用性场景进行了分析，发现这些方法在数据分布存在差异的情况下很难有效，然后进一步回顾了现在跨域泛化方法的优点和不足。紧接着根据上述分析和临床需求，提出了一种新的学习场景，即利用已有源域模型和目标域上有限标注数据实现高效跨域适应，无源主动域适应。并提出双参考样本选择策略从目标域中选择域无关和域相关样本进行标注以提高标注样本的多样性缓解数据发布差异。进一步提出多阶段半监督学习方法利用有限主动标注样本和无标注的目标域数据进行模型微调，进而实现高效的跨域分割。实验结果表明该方法可以在极低的标注成本下实现准确的分割，多中心临床验证结果也表明了该方法的有效性，为基于有限标注数据的医学图像分割方法在跨中心临床应用上提供了新的思路。

第七章，总结与展望。对本研究内容进行了总结和对未来研究进行了展望。

为了降低数据标注成本和模型对数据的依赖，本研究利用有限标注数据进行模型学习，第三章用测试阶段的用户交互作为有限标注，第四章用涂鸦标注作为有限标注，第五章用少量样本标注作为有限标注，第六章用主动选择的少量样本标注作为有限标注。从数据标注类型（点、涂鸦、部分全标注和选择样本进行标注）来说，这四章方法所需数据标注形式层层递进；从标注成本、标注难易程度和临床性能表现来说这四章方法结果层层递进，逐步达到临床可用性能，因此这四章属于递进的逻辑关系。

第二章 医学图像分割基础与有限标注学习

为了更好的阅读和理解本研究的后续工作内容，本章首先介绍常见的医学图像分割方法，包括本研究涉及到的传统分割方法和最新基于深度学习的医学图像分割方法；然后介绍本工作主要涉及的基于有限标注的医学图像分割相关方法的理论基础和应用实践，最后介绍了与本工作相关且被广泛使用的开源基准数据集。

2.1 医学图像分割基础

医学图像分割是指从医学图像中准确提取出对诊断、治疗或后续研究有帮助的区域，比如病灶、器官或组织等。作为计算机辅助诊断系统的重要组成部分，图像分割在器官组织定量评估、术前规划、疗效评估、预后分析等方面被广泛应用。常见的医学图像分割包含了 X 光图像、超声图像、眼底图像、病理图像等二维图像，以及 CT、MRI、PET、OCT 等三维断层成像。在图像分割的发展过程中，分割问题通常被定义为曲线的演化过程或像素的分类问题^[53,79,109,110]，并利用数学、优化、机器学习等工具进行求解获得分割结果。如1.2.1节所知，从方法学的角度来说图像分割大致可以被分为四类：1) 基于规则；2) 基于模型；3) 基于优化；4) 基于学习。随着人工智能的快速发展基于规则、模型、优化的图像分割方法也被统称为传统分割方法；基于学习的分割方法也主要发展为基于深度学习的分割方法。考虑到图像分割已经进入到了深度学习时代，本章主要介绍本研究涉及到的水平集和图割方法以及基于深度学习的分割方法，其余传统方法的相关内容介绍部分可以参考以下两篇综述论文^[110,111]。

2.1.1 水平集方法

水平集方法^[112,113] (Level set methods) 通过演化水平集函数来实现图像的区域分割，在实现过程中该演化过程通常被转换为偏微分方程进行迭代式求解^[53,60]。水平集函数 $\varphi(\mathbf{x})$ 隐含地表示了图像中不同区域的边界。以单类分割为例，水平集函数通常用如下带符号距离函数表示：

$$\varphi(\mathbf{x}) = \begin{cases} -\inf_{y \in \partial S} \|\mathbf{x} - \mathbf{y}\|_2, & \mathbf{x} \in S_{\text{in}} \\ 0, & \mathbf{x} \in \partial S \\ +\inf_{y \in \partial S} \|\mathbf{x} - \mathbf{y}\|_2, & \mathbf{x} \in S_{\text{out}} \end{cases} \quad (2-1)$$

x, y 分别表示图像中两个不同像素的位置坐标, ∂S 是零水平集, 也代表目标的轮廓。 S_{in} 和 S_{out} 分别表示分割目标的内部区域和外部区域。水平集函数 $\varphi(\mathbf{x})$ 通过演化方程来不断更新, 以寻找目标的轮廓。常见的水平集演化方程如下:

$$\frac{\partial \varphi}{\partial t} = F|\nabla \varphi| \quad (2-2)$$

其中 $\nabla \varphi$ 为水平集函数的梯度; $|\nabla \varphi|$ 为其梯度的模长; F 是演化速度函数, 控制水平集函数的演化方向和速度。通过调整演化速度函数 F , 可以实现对不同区域的边界的追踪和分割。后续的许多研究在原有方法的基础上引入了距离场来提升数值稳定性和对噪声及初始条件的鲁棒性^[114]; 或者是将其与变分原理相结合, 通过最小化能量函数来实现图像分割, 提高分割的全局一致性和对噪声的鲁棒性^[60]; 亦或扩展其允许分割多个相互交叉或重叠的区域, 从而更适用于复杂的图像分割任务^[27, 115]。最近随着深度学习的快速发展, 将深度学习与水平集方法相结合, 通过神经网络学习水平集函数的演化过程, 可以取得更好的分割结果和更快的收敛速度, 为后续的进一步探索提供了经验和指导^[116]。

2.1.2 图割法

图割 (GC) 是一种有效的基于能量优化的求解算法, 并被广泛应用于图像分割任务中^[2]。在 GC 中, 图像被表示为一个图, 其中图的节点表示图像中的像素, 边表示相连像素之间的关系, 如图2-1所示, 其中 s 被称为源点 (source) 连接像素为前景, t 被称为汇点 (sink) 连接像素为背景。不同于普通图, GC 有两个终端顶点 s 和 t 且图中所有节点都必须与终端顶点分别相连会形成两种类型的边: 1) 像素顶点和相邻像素的边, 也被成为 **n-links**; 2) 终端顶点 s 和 t 与所有像素点的连接, **t-links**。根据像素之间的相似性、空间位置等属性进行建模为每条边定义一个权值也被成为代价 (cost)。图2-1中割 (cut) 即为图中边集合的一个子集, 边子集的代价即为所有边的权值和。其中如果一个割, 它的边子集的代价最小, 则其称为最小割 (min-cut), 即图割的结果。根据图割的结果图像所有像素的顶点连接也被分为了 s 和 t 两个子集, 即与 s 相连的前景和与 t 相连的背景。基于此理论后续许多图像分割算法利用了图割作为基本工具并根据实际需求进一步优化^[24, 62]。本研究也进一步利用图割进行分割结果优化以处理未见过器官。

2.1.3 深度学习基础

计算机视觉旨在构建计算机系统来理解、解释和处理图像或视频数据, 其主要任务为图像分类、分割和检测。随着 AlexNet^[3] 在 2012 年 ImageNet^[30] 图像分类竞赛中取得的突破性进展, 深度学习在计算机视觉和图像处理领域获得了广泛

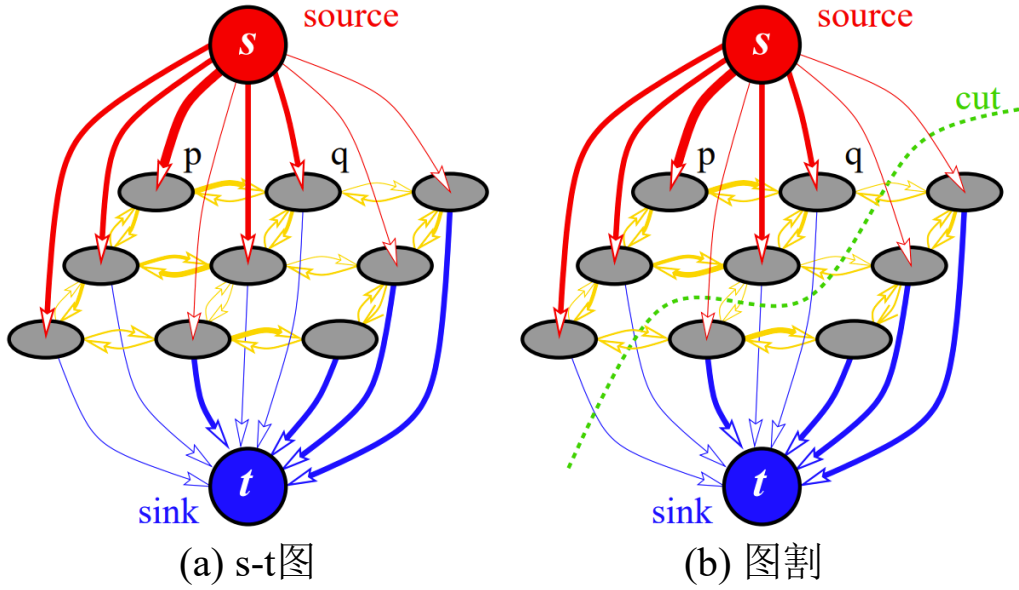


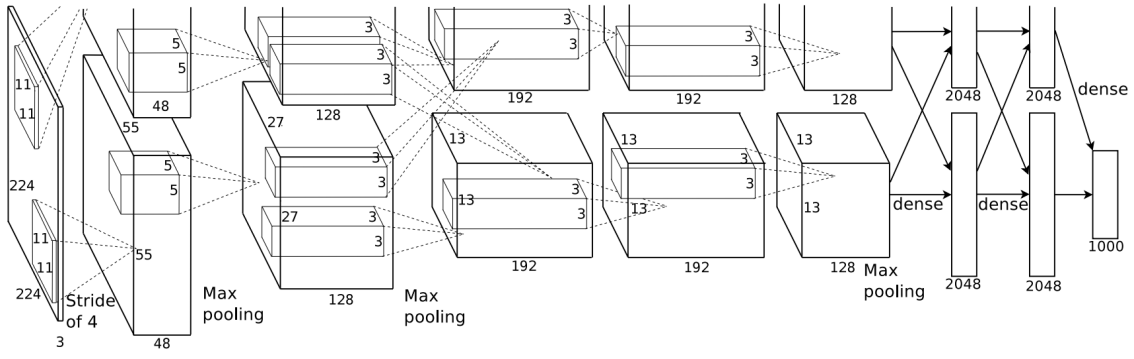
图 2-1 图割法中图的建立和图割示意图，图片参考自 [2]

关注。之后几年，随着 GoogleNet^[117]、VGGNet^[118]、ResNet^[29] 的相继出现，计算机视觉获得了空前发展，并首次在图像识别任务上准确率超过了人类水平^[119]，至此计算机视觉进入了深度学习时代。最近基于自注意力机制的 Transformer^[120] 架构在计算机视觉任务上也取得了突破性的进展^[67]，考虑到本研究主要使用卷积神经网络架构，因此本节主要介绍卷积神经网络的基础知识，Transformer 的相关内容可见最新技术和综述文章^[67, 120, 121]。

经典的深度学习方法主要由卷积神经网络和模型训练方法组成。卷积神经网络通过简单的卷积操作将原始输入数据逐步转换为高层次的语义特征，进而实现数据的特征表示。以经典的卷积神经网络结构 AlexNet^[3] 为例，如图2-2所示，可以发现其主要由输入层、卷积层、激活函数层、池化层、全连接层和输出层组成。对于二维卷积神经网络来说，输入层通常为原始图像的三维张量，即图像的高度、宽度和通道数（如彩色图像的 RGB 三通道，灰色图像的灰度通道等）。卷积层通过滤波器（卷积核）对输入数据进行卷积操作得到特征图，进而提取输入数据的空间结构信息。图2-3展示了一个卷积操作的示意图，以 5×5 图像，1 步长的 3×3 卷积核为例，卷积的计算过程可表示为：

$$a_{i,j} = f\left(\sum_{m=0}^2 \sum_{n=0}^2 w_{m,n} x_{i+m,j+n} + w_b\right) \quad (2-3)$$

其中 $x_{i,j}$ 表示图像第 i 行第 j 列元素、 $w_{m,n}$ 表示卷积核第 m 行第 n 列的权重、 w_b 表示卷积核的偏置项、 $a_{i,j}$ 表示卷积后得到的特征的第 i 行第 j 列元素、 f 表示激活函数。此外，根据输入数据的尺寸，卷积操作也会采用在输入四周填充零的操作来

图 2-2 卷积神经网络结构示意图，以 AlexNet^[3] 为例

调整输入特征以满足卷积计算需要。

从公式2-3可知，卷积操作是一种线性组合，为了提高神经网络的表达能力使其能够更好地适应和拟合复杂的数据分布，需要引入非线性变换（即激活函数）。目前常见的激活函数主要有以下四种：

1) Sigmoid 函数，将输入数据映射到 0 到 1 这个较小的范围，有助于模型训练，其数学表达式如下：

$$f(x) = \frac{1}{1 + e^{-x}} \quad (2-4)$$

如图2-4左上所示，Sigmoid 函数曲线为 S 型，其数值随着输入的增大变化越来越小，因此存在饱和效应，即输入较大或较小时 Sigmoid 函数激活会导致梯度消失，导致模型训练崩溃。

2) Tanh 函数，将输入映射到-1 到 1 这一范围内，相比 Sigmoid 函数具有更强的非线性能力；且其以 0 为中心的特性，有利于缓解梯度消失等问题，其数学表达式如下：

$$f(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} \quad (2-5)$$

如图2-4右上所示，Tanh 函数曲线与 Sigmoid 曲线类似，也存在梯度消失的问题。

3) ReLU 函数，对于正数输入，返回输入值，对于负数输入，返回零，其数学表达式如下：

$$f(x) = \max(0, x) \quad (2-6)$$

ReLU 函数曲线如图2-4左下所示，相比 Sigmoid 和 Tanh 函数，ReLU 函数计算更简单、收敛速度更快。此外 ReLU 函数能在提高网络的稀疏性的同时减少参数量，进而加速了模型训练。但是当输入小于 0 时，其梯度为 0，会导致该部分神经元变

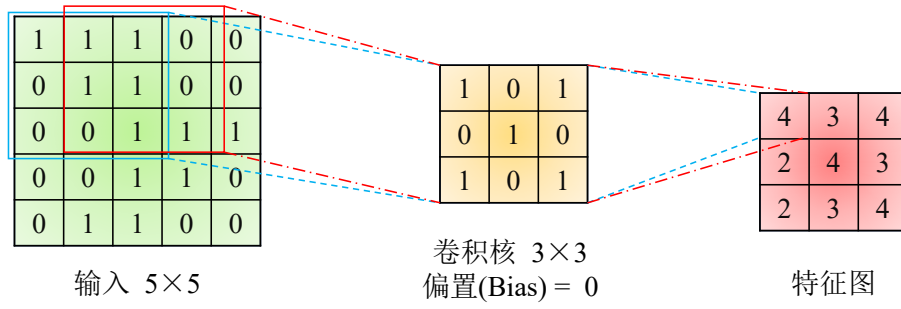


图 2-3 卷积操作示意图，以 5×5 图像和 3×3 卷积核和 1 步长为例

得不活跃（即神经元“死亡”现象），不利于模型训练。

4) Leaky ReLU 函数，引入了带参数 (a) 的表示方法缓解当输入小于 0 时 ReLU 函数存在的神经元“死亡”现象，其数学表示如下：

$$f(x) = \begin{cases} x, & x \geq 0 \\ ax, & x < 0 \end{cases} \quad (2-7)$$

图2-4右下展示了 Leaky ReLU 函数的激活曲线。与 ReLU 函数相比，Leaky ReLU 的形式略微复杂（引入了一个 a 参数，通常设置为很小的常数，如 0.01），但仍保留了 ReLU 收敛速度快和计算简单等特点。鉴于 Leaky ReLU 函数的上述特点，其成为了最流行和广泛使用的激活函数。为了在降低特征尺寸和参数量的同时保持特征信息，池化操作被广泛应用于卷积神经网络训练过程中。池化层通常位于卷积层之后，对卷积得到的特征进行下采样，通过简单的统计计算即可完成。最常见的池化操作有两种，最大池化和平均池化，计算过程如图2-5所示。在最大池化中，池化窗口内特征的最大数值被保留，而其他值被丢弃，这种池化方式有助于保留特征中的最显著的部分，降低特征尺寸并提高模型的计算效率。平均池化是指用池化窗口内特征的平均值作为池化后的特征。与最大池化不同，平均池化考虑了池化窗口内所有特征信息，而不仅仅是最大值，因此可能会弱化一些显著特征。根据它们各自特点，通常被应用于不同的任务中，例如分类任务为了捕捉图像显著特征经常使用最大池化，而分割任务为了获得平滑的预测结果通常使用平均池化。此外，基于整个特征图的全局池化操作也常常被应用于特征压缩等任务中^[122]。

全连接层将卷积或池化后的特征用于分类或回归。具体来说，全连接层首先将特征转换为一维向量，然后采用密集连接的方式构建不同特征之间的相互关系，得到高层次的特征向量，用于输出层预测。全连接及输出层的计算过程如下：

$$y = f\left(\sum_{i=0}^n w_i x_i + w_b\right) \quad (2-8)$$

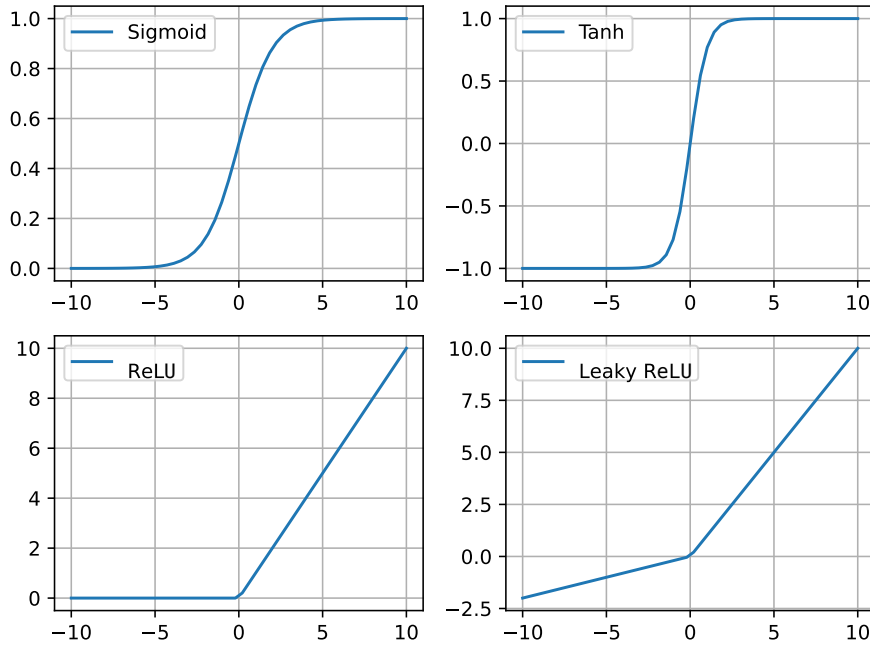


图 2-4 四种不同激活函数示意图

其中 f 为输出层，通常为激活函数； y 表示全连接层的输出结果； x_i 为一维特征； w_i 和 w_b 分别为全连接层的模型参数和偏置。此外，如何训练卷积神经网络也是深度学习的研究重点，其主要包括损失函数和优化算法。损失函数不仅被用于衡量模型性能和指导参数更新，还定义了模型的优化目标，反映了任务的特性，并且有助于防止过拟合。常用的损失函数为交叉熵函数（Cross-Entropy Loss, CE），以二分类为例，其定义如下：

$$L_{CE} = -\frac{1}{n} \sum_{i=1}^n [y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)] \quad (2-9)$$

其中 n 是样本数量； y_i 是真实的类别标签（数值为 0 或者是 1）； $\hat{y}_i$ 为模型预测的类别概率（属于 0 到 1 之间）。与传统方法类似，深度学习也采用反向传播方法（Back Propagation, BP）和梯度下降法（Gradient Descent, GD）通过最小化损失函数来迭代的更新模型参数，迫使模型从标注数据中进行自动学习。目前，常有的梯度下降方法包括随机梯度下降^[123]（Stochastic Gradient Descent, SGD）和其改进后的自适应矩估计（Adaptive Moment Estimation, ADAM）^[124]。此后深度学习研究主要集中于网络结构设计和模型训练方法探索。

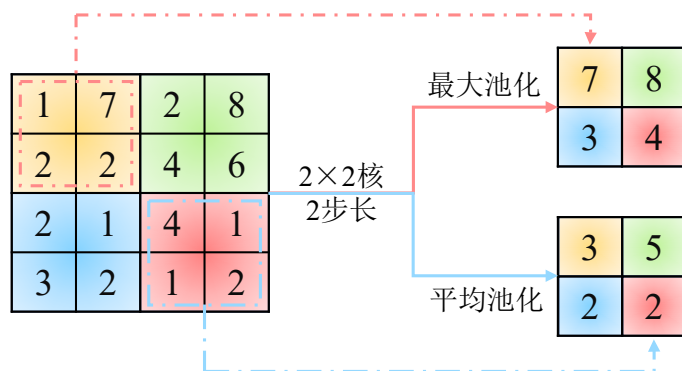


图 2-5 池化操作示意图

2.1.4 基于深度学习的图像分割

受到图像识别任务的启发，Long 等人^[31]首次利用全卷积神经网络（FCN）开发了基于深度学习的全自动端到端图像分割方法，并在多个数据集上取得了最佳分割性能。为了实现图像分割，FCN 在分类网络 VGGNet^[118]的基础上进行了三点主要改进。首先是用卷积层代替 VGGNet 中的全连接层，从而使得网络可以接受任意尺寸的输入图像，并输出相同尺寸的预测结果，实现了端到端的像素级别分割。其次是引入了反卷积操作来将特征图逐步上采样到输入图像的尺寸，使得网络可以为每个像素生成预测结果。最后是引入了跳跃链接，通过跳跃链接网络可以把不同层次的特征相结合，同时保留细节信息和上下文信息，提高模型的分割性能。

之后，基于深度学习的图像分割方法得到了快速发展，图2-6展示了近几年提出的有较大影响力的分割方法。例如，通过引入对称的编码器-解码器的网络架构和改进跳跃链接结构将编码器中的低级和高级特征图连接到解码器中的相应层为医学图像分割设计了新的网络结构，U 型网络^[8]（U-Net）。相比 FCN，U-Net 能保留更多的空间信息和细节信息进行更精细的图像分割。得益于这种简单有效的网络结构，U-Net 在众多医学图像分割任务上取得了优异的结果^[16]，并成为了医学图像分割任务中应用最广、最有效和最有影响力的方法^①。随后 U-Net 被大量改进和扩展用于解决特定分割任务^[5,32,125-128]，包括二维扩展版本 U-Net++^[33]及其不同的三维扩展版本，3D U-Net^[32]、V-Net^[5]。本研究中主要使用 U-Net^[8]、3D U-Net^[32]、V-Net^[5] 作为基线模型。

继 FCN 之后，DeepLab 系列分割方法^[129-131]再度将图像分割推上了一个新的高度。针对已有工作存在的下采样导致细节信息丢失和 CNN 空间不变性导致的难以捕获细节信息，DeepLabV1^[129]提出了空洞卷积（Atrous Convolution）在不降低

① 据谷歌学术显示 U-Net^[8] 提出至今已引起超 80000 次，是引用最高的图像分割方法。

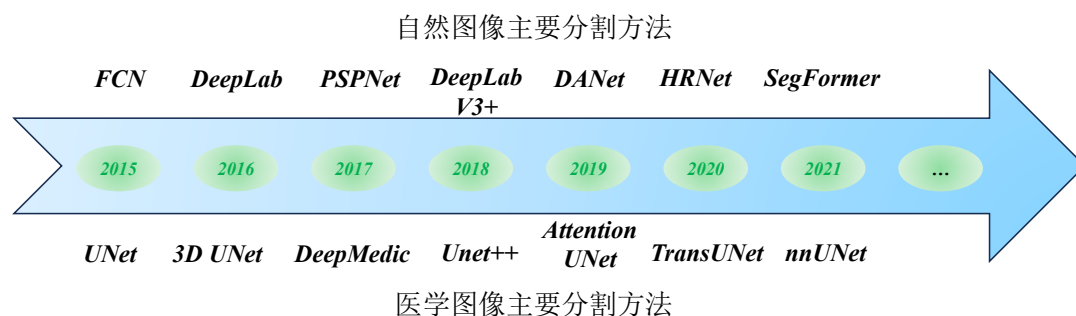


图 2-6 基于深度学习的图像分割方法发展历程

特征图分辨率的情况下增加模型的感受野进而更好的捕获细节信息，同时使用条件随机场^[22]来对模型预测进行修正。在此基础上，DeepLabV2 和 DeepLabV3^[130]提出了空洞空间金字塔（Atrous Spatial Pyramid Pooling, ASPP）通过使用不同空洞率的空洞卷积来提取不同大小物体的语义信息，并引入全局池化添加全局上下文信息来捕获细节信息，并在模型训练策略上使用了多项式衰减策略来调整学习率。之后，DeepLabV3+ 在网络结构上引入了类 U-Net 的编码器-解码器结构，并用深度可分离卷积（Depthwise Separable Convolutions）代替标准卷积，并不再用条件随机场来做后处理。与此同时，PSPNet^[132]提出了金字塔池化模块来整合全局上下文信息并利用深度监督加速模型训练^[133]。这些方法在自然图像上的成功也启发医学图像分割方法的研究，比如 DeepMedic^[134]将 FCN 从二维扩展到三维用于三维医学图像分割，并利用三维条件随机场进行预测优化，从而取得 2016 年脑卒中分割比赛的第一名^[135]。

之后，注意力机制以其能根据模型输入对特征图的不同空间、区域给予不同的关注度被广泛应用于深度学习中^[120,136,137]。DANet^[138]通过引入通道注意力和空间注意力来自适应的关注全局特征和局部依赖，并在多个数据集上超越了 DeepLabV3+ 等最新方法。在医学图像上，Attention-UNet^[139]首次将自适应空间注意力与 U-Net^[8]和 3D U-Net^[32]相结合，设计出符合医学图像特点的分割网络。随着深度学习基础模型的快速发展，高分辨率网络和 Transformer 结构被广泛应用于图像分割任务中，诞生了一些列高性能图像分割算法，例如 HRNet^[140]、TransUNet^[141]、SegFormer^[142]等。尽管每年有众多图像分割网络被相继提出，从结构上来看它们依旧是 U-Net^[8]或 FCN^[31]的衍生结构。最近，Fabian 等人在原有 U-Net^[8]和 3D U-Net^[32]的基础上提出了 nnUNet^[34]（No new U-Net），该结构在大量医学图像分割任务上取得了最佳的分割性能，并以其大道至简的设计思路为医学图像分割提供了可靠基线。大量数据集上的实验结果表明，针对特定任务进行人为的模型结构调整可能会导致模型过拟合，数据处理和训练策略对分割的影响

更大。至此，医学图像分割任务的主要研究方向从网络结构设计走向了训练策略及通用性研究。

2.2 高效标注与有限标注学习

标注数据有限是医学图像分割任务中普遍存在的问题，主要由于医学图像标注繁琐耗时且依赖临床经验^[70]，此外隐私保护也导致医学图像及标注难以共享。然而深度学习是数据驱动的方法，其优异的表现依赖大量高质量标注数据，比如训练 ResNet^[29]、FCN^[31] 这种深层分类、分割网络，需要大量的标注数据（常见的 ImageNet 分类数据集包含了约 150 万张带标注的图片，ADE20K^[143] 分割数据集包含了约 3 万张像素级标注图像）。为了降低深度学习网络对标注数据的依赖，探索如何利用有限标注数据来开发高性能模型获得了计算机视觉及医学图像处理领域的广泛关注，并取得了长足的进步^[95,97,144-146]。图2-7展示了目前有限标注学习的主要研究问题及方法、其分类标准参考自^[70]，下面将详细介绍相关方法的技术基础。

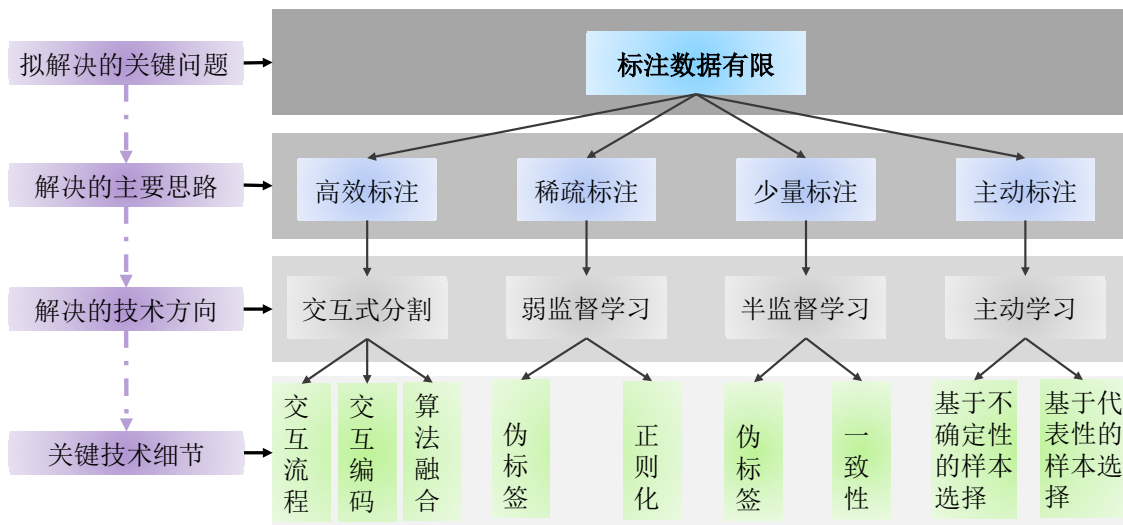


图 2-7 有限标注学习的主要研究问题及方法

2.2.1 交互式分割

交互式分割即利用用户交互来指导算法进行图像分割和标注。为了实现用户与算法的有机组合，有三个主要问题亟待解决，首先是交互流程，即什么交互形式和如何交互；其次是如何编码这些交互信息使其成够有效指导算法进行准确预测；最后是如何将这些编码信息与模型相融合进行结果修正。

交互流程： Xu 等人^[81] 首次提出利用点和全卷积网络 FCN 来进行图像的交互

式分割，该流程要求用户为每一个分割目标提供一个前景点和背景点来指导深度学习模型准确分割选定目标。同时，Wang 等人^[77]提出了一种两阶段交互式分割方法，首先第一阶段利用自动分割模型来预测初步分割结果，然后用户根据初步分割结果用涂鸦指导修正模型进行结果优化。尽管上述两种交互式流程被大量后续的交互式分割方法采用，其依旧难以处理未见过的目标。之后，Wang 等人^[76]用边界框作为初始交互来指导深度学习模型分割边界框内的器官，然后利用涂鸦交互和条件随机场约束^[22]针对每张测试图像进行模型微调，进而实现见过和未见过器官的交互式分割。但目前这些交互式分割方法只能处理单一模态图像和器官，难以进行复杂场景下的图像分割和标注。

交互编码：在交互式分割中用户交互信息（点、线、框）往往较为稀疏，如何将其转换为深度学习模型可以充分利用的输入形式也是研究的重点。Xu 等人^[81]提出用欧式距离变化（Euclidean Distance Transformation, Eucl）来将稀疏的点击转变为密集的距离信息，其转换形式如下：

$$Eucl(s) = \min_{(i_x, i_y) \in I} \sqrt{(i_x - s_x)^2 + (i_y - s_y)^2} \quad (2-10)$$

其中 (i_x, i_y) 为图像 I 中像素 i 的 x, y 的坐标信息； s 为用户交互点， (s_x, s_y) 为 s 的坐标信息。 $Eucl(s)$ 为与图像尺寸大小一样的欧式距离图，每个像素的数值为该点到交互点的最小欧式距离。之后，Maninis 等人^[82]在欧式距离的基础上引入了高斯核函数来编码用户交互信息（Gaussian Transformation, Gauss），其数学定义如下：

$$Gauss(s) = e^{-\frac{\min_{(i_x, i_y) \in I} (i_x - s_x)^2 + (i_y - s_y)^2}{2\sigma^2}} \quad (2-11)$$

其中 σ 控制了用户交互信息的影响范围，其值越大，影响范围越大。与欧式距离变换相比，高斯变换能更好的强化交互点附近的结果，避免远离交互区域的错误更新。此外，Wang 等人^[76]首次将语义编码能力更强的测地线距离变换^[71]（Geodesic Distance Transformation, Geos）用于交互点信息编码，其变换形式如下：

$$Geos(s) = \min_{p \in \mathcal{P}_{s,I}} \int_0^1 \|\nabla I(p(n)) \cdot \mathbf{v}(n)\| dn \quad (2-12)$$

其中， $\mathcal{P}_{s,I}$ 是用户交互点 s 和图像 I 中其他像素之间的所有路径的集合。 p 是一条可行路径，并且由 $n \in [0, 1]$ 参数化。 $\mathbf{v}(n) = p'(n) / \|p'(n)\|$ 是沿着路径方向的单位向量。与欧式距离变换和高斯变换相比，测地线变换能更好的突出交互点区域与其他区域的语义差异。

人工交互与模型融合：此外，探索如何结合用户交互信息和算法模型来提高分割性能也是研究的重点。Xu 等人^[81]和 Wang 等人^[77]直接在输入层面将交互信

息编码图和图像拼接送进深度学习模型进行自动学习和推理，鼓励模型学习交互信息和分割结果之间的相互关系。Hu 等人^[74]利用双编码器模型先分别提取交互信息编码图和原始图像的高阶特征，然后在特征层面进行融合进行交互式分割。这两种策略都是只从输入的层面考虑了交互信息，没有直接建模交互信息与模型输出的关系。为了改善上述缺陷，Wang 等人^[76]在其既往工作^[77]基础上，引入了条件随机场约束^[22]将交互信息融入损失函数中来指导模型微调，直接建模了交互信息和输出结果的关系，能更好的提高交互式分割的准确性。但如何利用交互信息提高模型对未见过图像的分割能力方面依旧缺乏研究。

2.2.2 弱监督学习

作为有限标注的一种，稀疏标注也被广泛用于训练深度学习模型，降低标注成本的同时提高模型的分割性能。不同于全监督学习，图像中的每个像素都有标注作为监督信息；稀疏标注只提供极少量的监督信号，如图像的类别信息，或极少数像素的标注信息，如图1-4所示。针对稀疏标注缺少监督信号的问题，目前主要有以下两种解决方案：

基于伪标签的弱监督图像分割：伪标签利用已有方法或模型为未标注的像素生成预测，并使用这些预测结果作为监督信号进行模型训练。Huang 等人^[147]将类别激活图（Class Activation Mapping, CAM）与图像标签相结合提出了基于类别信息的弱监督分割方法。该方法首先利用图像标签来训练分类模型，然后利用训练好的分类模型和 CAM 技术提取图像中特定类别的激活图，之后用阈值和条件随机场来修正激活图生成伪标签用于分割模型训练。此外，针对涂鸦标注问题，Lin 等人^[87]提出了基于图模型的^[148]标签传播方法，根据特征相似性把涂鸦标注传播到未标注的像素生成伪标签，然后利用伪标签训练模型。Can 等人^[88]把涂鸦标注作为输入种子点利用 Random Walks^[72]生成伪标签，来训练深度学习模型，并进一步利用条件随机场^[22]修正模型预测再次生成伪标签训练模型，迭代的更新伪标签和训练模型。后续，基于不确定性修正被相继提出用于筛选高质量伪标签，为弱监督学习提供了新的伪标签生成方法^[91]。

基于正则化的弱监督图像分割：此外，基于数学统计的正则化方法也被广泛用于弱监督学习^[7,149]。Tang 等人^[7]首次利用马尔可夫模型来建模像素之间的空间关系并将其定义为空间约束损失函数，数学形式如下：

$$L_{spatial} = - \underbrace{\sum_{i \in \omega_s} \log y_i}_{\text{第一项}} + \lambda \times \underbrace{\sum_{i,j} w_{i,j} \times |y_i - y_j|^2}_{\text{第二项}} \quad (2-13)$$

其中 i 和 j 分别代表图像中的两个像素点, y_i 和 y_j 分别表示它们的预测结果。 w_{ij} 是像素 i 和 j 之间的权重, 通常根据它们之间的空间距离和像素之间的相似度来计算, 如高斯函数或者指数函数^[149]。第一项是利用标注像素信息进行监督学习的过程, 该过程只计算有标注像素的损失。第二项是指空间和像素层面的正则化项, 利用这些空间和像素统计信息来约束模型从未标注像素上学习。不同于伪标签策略, 正则化方法不需要多次标签迭代更新, 直接以端到端的形式学习, 此外相比于全监督学习, 正则化方法只引入了一个额外的正则化损失函数, 在训练成本上也没有额外开销。之后, 正则化损失成为了弱监督图像分割研究的基础模块。

2.2.3 半监督学习

半监督学习利用有限标注数据和大量未标注数据进行模型训练, 减少标注成本的同时可以有效提升模型分割性能, 近年来其也逐步成为了研究热点^[70,150]。按监督信号来源分类, 目前半监督学习大致可以分为伪标签和一致性两类, 如图2-8所示。

伪标签: 与弱监督学习类似, 伪标签方法也被大量应用于半监督学习中^[93,146]。David 等人^[145] 提出了 MixMatch 方法, 通过对有标签数据 and 无标签数据进行混和增强来生成伪标签并利用伪标签来进行模型训练, 如图2-8(a) 上所示。MixMatch 对输入的无标签图像进行一系列增强, 然后送进模型进行预测, 并根据这些预测的平均值利用标签平滑策略生成伪标签, 然后用伪标签来进行监督学习并更新模型参数。Kihyuk 等人^[144] 在 MixMatch 的基础上提出了 FixMatch, 如图2-8(a) 中所示。FixMatch 通过对未标注图像进行两种不同的数据增强、弱增强 (weak augmentation, x_w) 和强增强 (strong augmentation, x_s), 然后利用弱增强图像预测生成伪标签来监督强增强图像预测。为了提高模型的性能, 在生成伪标签时, FixMatch 只考虑高置信度的预测用于模型训练。同时, PseudoSeg^[151] 将 FixMatch 应用于图像分割任务中并在多个分割数据集上取得了最好的分割结果。以上两种伪标签策略主要是从数据层面引入了数据增强来生成多样性的预测并进一步生成用于模型训练的伪标签。Chen 等人^[11] 尝试在模型层面引入了差异来生成互补的伪标签进行相互监督 (Cross Pseudo Supervision, CPS), 其方法如图2-8(a) 下所示。不同于 MixMatch、FixMatch, CPS 采用了两个分割网络来执行图像分割, 这两个分割网络有着一样的模型架构但初始化参数不同, CPS 使用每个模型的输出直接生成伪标签用于另一个模型训练, 该过程保留了每个模型训练的独立性, 有利于产生更多多样性的伪标签进行互补学习。此外, 一些研究也通过引入了不确定性修正^[152,153] (Uncertainty Rectification)、不同网络结构^[153]、条件随机场^[95,154] 等

方法来提高伪标签的质量。

一致性：除了伪标签外，一致性学习也被广泛用于半监督学习。Antti 等人^[94]提出了一个新的半监督学习策略，教师-学生模型（Mean Teacher），通过鼓励学生模型输出和教师模型一致来利用无标注数据进行学习，如图2-8(b)上所示。具体而言，该模型由网络架构一样的教师模型和学生模型组成，在训练过程中学生模型通过有标注数据的监督损失和一致性损失进行模型参数更新，教师模型参数则是通过学生模型参数的指数移动平均（Exponential Moving Average, EMA）计算得来，其中一致性损失为教师和学生预测概率的均方误差（Mean-Square Error, MSE）。尽管最新很多半监督学习方法被相继提出，教师-学生模型依旧是最流行的方法之一，例如在医学图像任务上 Yu 等人^[97]引入了不确定性修正^[152]、Li 等人^[155]引入了数据增强^[144]来改进教师-学生模型在多个医学图像分割任务上表现优异。此外，基于多个模型的协同一致性也被广泛用于半监督医学图像分割任务中^[156,157]，其核心思想是利用两个或多个模型来构造相互之间的一致性为无标注数据提供监督信息，其学习流程如图2-8(b)中所示。最近，基于模型不同分支的一致性学习也获得了研究者的广泛关注，其核心思路如图2-8(b)下所示。例如，Yassine 等人^[158]通过给 PSPNet 添加了多个辅助分割分支，并在主分支和辅助分支之间构建一致性，设计了新的半监督图像分割方法。与之类似，Wu 等人^[10]也证明了不同分割分支之间的相互一致性（Mutual Consistency）可以提高模型对无标注数据的利用效率，并在多个半监督医学图像分割任务上取得了性能提升。目前这些基于伪标签和数据、模型一致性的半监督学习方法只从输入和模型的角度来构建监督信息，没有考虑任务本身的特点，因此后续研究可以结合任务本身几何或解剖学特点进行监督信号设计。

2.2.4 主动学习

为了进一步在标注成本有限的情况下提高模型的分割性能，主动学习被提出用于根据标注成本需求来主动选择最有价值的样本进行标注。目前主动学习在多个领域获得了广泛的关注，并被应用于多个计算机视觉及医学图像处理等实际场景中^[1,70,159]。图1-5展示了经典主动学习的算法流程，可以发现主动学习的研究重点是如何选择有高信息量的样本用于标注和模型训练。最近 Budd 等人^[1]系统地回顾了主动学习的发展历程，并将样本选择方法主要分为不确定性和代表性两类。

基于不确定性的样本选择：目前不确定性是主要的样本信息量度量方法，其主要基于预测不确定性高的样本可以为模型训练提供更多信息这一假设^[160,161]。目前计算不确定性的方法主要有三种。以分割任务为例，最简单的不确定性计算

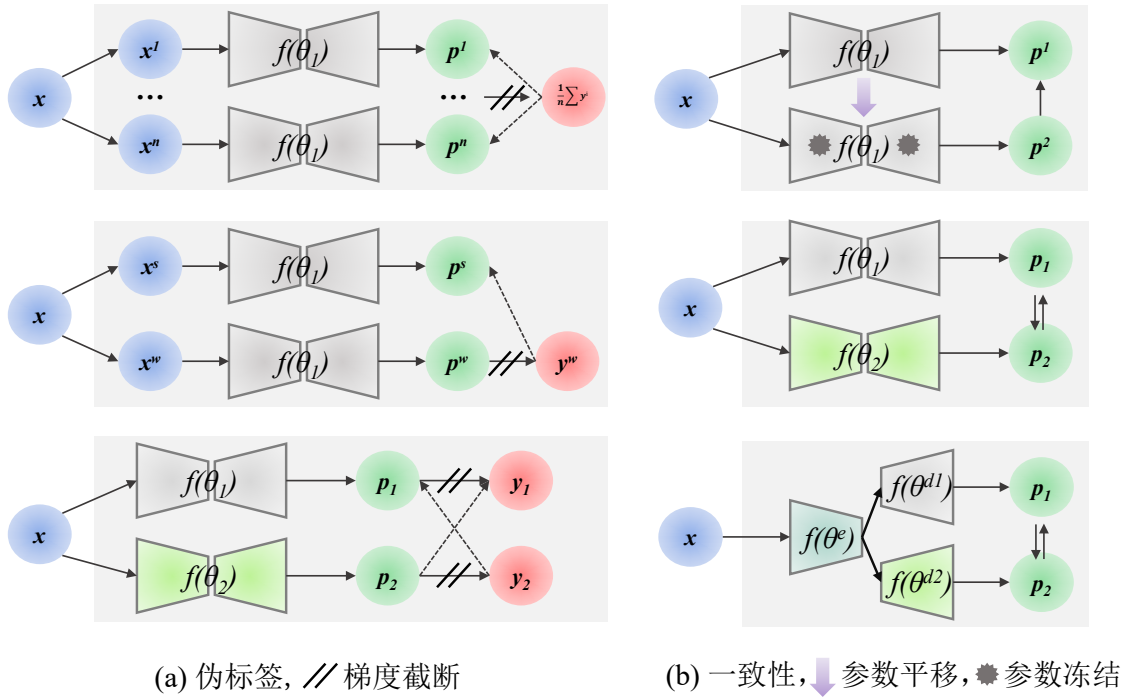


图 2-8 不同半监督学习策略示意图

方法是计算预测结果中每个像素的最低类别的概率之和^[162]。从直觉上讲更确定的预测往往有更高的预测概率，因此图像中每个像素的最低类别概率之和越高，则该预测被视为越不确定，其中每张图像的不确定性计算如下：

$$x_{LC} = \sum P_{\theta}(\hat{y}|x) \quad (2-14)$$

其中 $\hat{y} = \operatorname{argmin}_y P_{\theta}(y|x)$ ， $\hat{y}$ 表示每个像素预测概率最低的类别。 x_{LC} 数值表示全部像素最低类别概率之和。这是一种非常简单的不确定计算方式，并可以根据每个图像的 x_{LC} 来对整个数据集进行排序，并选取 x_{LC} 最大的前 n 个样本进行标注，这种方法被称为最小置信采样^[162]（Least Confident Sampling, LC）。它也存在着缺陷，因为它只考虑了最不确定的图像而没有考虑其他数据分布信息。为了缓解这种问题，两种不同方法被相继提出，首先是被称为边界采样（Margin Sampling）。边界采样在多分类场景下同时考虑第一个和第二个最可能类别预测差异（Best-versus-Second Best, BvSB）作为不确定性。这种方法的基本思想是两个最有可能的类别预测之间的间隔越大，模型预测该类别的置信度越高，其计算方式为：

$$x_M = \sum (P_{\theta}(\hat{y}_1|x) - P_{\theta}(\hat{y}_2|x)) \quad (2-15)$$

其中 $\hat{y}_1$ 、 $\hat{y}_2$ 分别表示每个像素预测概率最高和第二高的类别。第二种流行的不确

定性估计方法为香农熵^[163] (Shannon Entropy), 其数学表达式为:

$$x_E = - \sum P_{\theta}(y|x) \log P_{\theta}(y|x) \quad (2-16)$$

熵被广泛用于衡量编码一个数据分布所需的信息量, 因此熵通常被视为不确定性度量的重要手段。以二分类为例, 模型预测结果概率越靠近 0 或 1 时, 通常认为不确定性较低; 相反当模型预测概率在 0.5 附近时, 不确定性较高。以上三种不确定性估计方法均符合该结论, 但香农熵以其简单有效的表达形式被广泛应用于各种主动学习场景中^[162]。

基于代表性的样本选择: 尽管基于不确定性的样本选择策略取得了不错的结果, 但其仅关注不确定性可能导致选取的样本只聚焦于整个数据分布中很小区域(预测不确定区域), 并且仅训练来自分布相同区域的样本会导致标注样本中出现冗余, 并导致标注成本浪费。为此, 代表性样本选择策略 (Representativeness) 被提出用于从不同的数据分布区域进行采样, 增加样本的多样性, 从而提高主动学习的性能。具有高代表性的样本覆盖了分布同一区域许多图像的信息, 因此无需从同一区域挑选除最具代表性图像外的其他的图像^[162]。Huang 等人^[106] 利用聚类方法^[164] 从整个数据集中选择每个簇中最靠近聚类中心的样本作为最具有代表性的样本。最近, Katerina 等人^[165] 利用对比学习提取特征, 然后进行特征聚类选择最具代表性的样本用于模型训练, 实现了冷启动的主动样本选取。目前, 将不确定性和代表性相结合用于样本选择也被广泛用于医学图像分析任务中^[1,70,159]。这些方法的成功基于所有样本属于同一数据发布这一假设, 针对不同中心、不同数据分布的域差异场景, 依旧缺乏研究和验证。

2.3 有效标注分割数据集

为了验证本文提出方法的有效性并与现有方法进行公平对比, 本研究主要使用了几个公开且常用的基准数据集来进行相关实验。下面将简要回顾本研究实验使用的基准数据集的基本信息。

2.3.1 多模态脑肿瘤分割 MRI 数据集

多模型脑肿瘤分割 MRI 数据集 (Brain Tumor Image Segmentation, BraTS) 是医学图像分割领域最有名的数据集之一^[166], 由宾夕法尼亚大学等团队从 2012 年开始组织举办, 也是医学图像分割领域最享有盛誉的比赛之一^①。其数据集由来自多个医疗中心的高分辨率多模态核磁共振成像 (MRI) 序列组成。这些序列包

① <https://www.med.upenn.edu/sbia/brats2018/data.html>

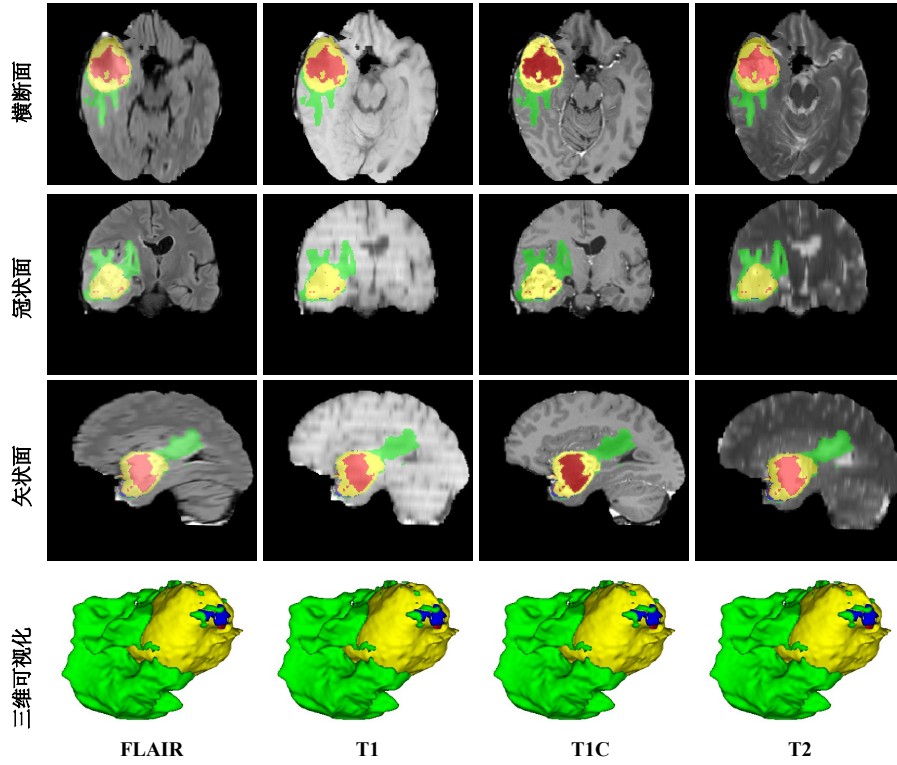


图 2-9 多模态脑肿瘤分割数据集

括 T1 加权 (T1-weight image, T1)、T1 增强 (T1-weighted contrast-enhanced image, T1C)、T2 加权 (T2-weighted image, T2) 和 T2 加权 FLAIR (T2-weighted FLAIR image, FLAIR), 这些序列都进行了配准, 可提供丰富的图像信息以用于肿瘤分割, 如图2-9所示。每个病人都由影像科医生进行了肿瘤分割, 包括肿瘤的核心区域 (包括增强和非增强部分)、水肿区域和囊变区域。除了自动分割, BraTS 也被广泛用于交互式分割的相关研究并逐步发展成为了基准数据集, 目前大量交互式医学图像分割方法都汇报了其在 BraTS 上的分割表现^[76, 77, 167, 168]。与自动分割不同, 这些方法主要集中于从 FLAIR 图像上分割整个肿瘤和从 T1C 图像上分割肿瘤核心的二分类任务^[166]。本文利用该数据集验证了提出的可泛化交互式分割标注方法的有效性。

2.3.2 左心房分割 MRI 数据集

左心房 (Left Atrium, LA) 分割 MRI 数据集是一个高质量的医学图像分割数据集, 由 AtriaSeg2018 比赛官方提供^[169], 其数据及标注形式如图2-10所示。该数据集由 60 名匿名房颤患者的 154 个独立获取的三维晚期钆增强心脏磁共振成像 (3D LGE-MRIs) 组成。其中三维晚期钆增强心脏磁共振成像的空间分辨率为 $0.625 \times 0.625 \times 0.625 \text{ mm}^3$, 其空间尺寸为 $576 \times 576 \times 88$ 或 $640 \times 640 \times 88$ 。LA 体素级别

的标注由三个训练有素的医生共同完成最后由一致共识标注作为分割金标准。由于其较大的数据规模和较高的标注质量，该数据集迅速成为了医学图像分割领域重要的基准数据集。尤其是在半监督医学图像分割领域，Yu 等人^[97]首次在该数据集上验证了 Mean Teacher 的有效性，为后续研究提供了公平的对比实验设置，推进了该数据集的广泛应用^[10,170,171]。本研究利用该数据集验证了提出的半监督分割方法的有效性。

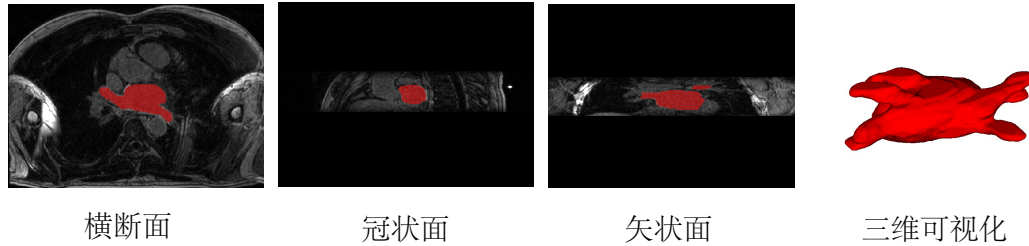


图 2-10 左心房分割数据集示例

表 2-1 本文使用到的多中心腹部多器官分割数据集

数据集	年份	器官	数量
BTCV	2015	脾脏、肾（左、右）、胆囊、食管、肝脏、胃、主动脉、下腔静脉、门静脉和脾静脉、胰腺、肾上腺（左、右）和十二指肠	50
TCIA	2018	脾脏、右肾、胆囊、食管、肝脏、胃、脾静脉、胰腺和十二指肠	90
MSD	2018	肝脏、脾脏和胰腺	682
WORD	2021	肝脏、脾脏、肾（左、右）、胃、胆囊、食管、胰肾上腺（左、右）左肾上腺、右肾上腺、直肠、膀胱和股骨头（左、右）	150

2.3.3 腹部多器官 CT 数据集

在医学图像分割任务中，腹部多器官数据集也被广泛用于算法验证。目前最广泛使用的腹部 CT 数据集为：BTCV^①、TCIA^[172]、MSD^[16]和 WORD^[44]，其详细统计信息如表2-1所示，标注信息如图2-11所示，其中 MSD 的标注来自 WORD^[44]工作。从表2-1可知、腹部多器官分割数据集主要集中于分割上腹部的肝脏、脾脏、胰腺等正常器官。既往其中数据集 TCIA 中包含了 BTCV 里的 43 个病人，其不同在于 TCIA 对 BTCV 的部分未标注测试数据进行了标注。此外、MSD 腹部数据集主要由三个子集组成，其中肝脏分割数据集包含 201 例三维 CT 图像及肝脏及肿瘤标注、脾脏分割数据集包含 61 例 CT 图像及脾脏标注、胰腺分割数据集包含 420 例 CT 图像及胰腺和肿瘤标注。WORD 数据集是较为全面的全腹部器官分割数据集，包含了 150 例全腹部 CT 图像及 16 个主要腹部器官标注。不同于其他数据集，WORD 数据集除了提供了像素级标注外还为每个图像提供了涂鸦标注用于弱监督学习研究，如图2-12所示。此外，TCIA 数据集中的胰腺分割任务也被广泛用于半监督学习研究^[157,173]。本研究利用该数据集分别验证了提出的可泛化交互式分割

① <https://www.synapse.org/Synapse:syn3193805>

标注方法和基于涂鸦分割方法的有效性。

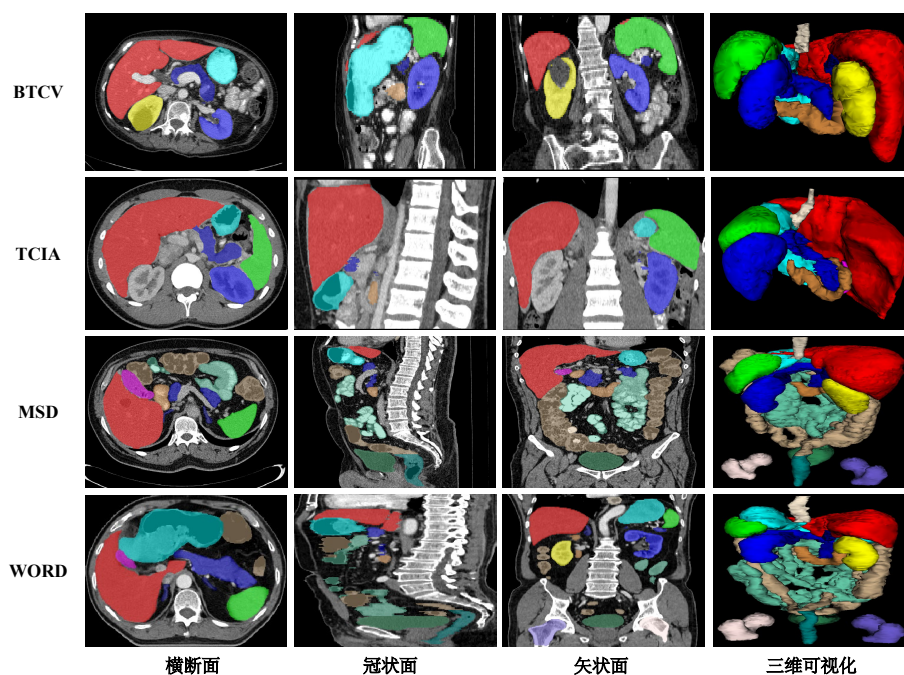


图 2-11 不同腹部多器官分割数据集示例

2.3.4 心脏结构分割 MRI 数据集

2018 年, Olivier 等人^[174]构建了一个基于 MRI 图像的心脏自动诊断和分割数据集^① (Automated Cardiac Diagnosis Challenge, ACDC) 用于算法研究。该数据集包含了 150 个患者的 300 个三维 MRI 图像 (每个病人分别有一个收缩期和舒张期图像), 这些患者主要包含了四个疾病组和一个健康的对比组。同时, 高年资的医生手工标注了每个 MRI 图像的三个器官: 左心室 (Left Ventricle, LV)、右心室 (Right Ventricle, RV) 和心肌 (Myocardium, Myo)。之后, 该数据集因其大规模和精细标注迅速被大量用于算法研究^[175]。最近, Valvano 等人^[92]基于该数据集探索了基于涂鸦的弱监督学习, 并迅速成为了医学图像弱监督分割的基准数据集。图2-13展示了心脏结构分割 MRI 数据集的像素级标注和涂鸦标注的差异。本研究利用该数据集验证了提出的基于动态混合伪标签的涂鸦监督方法的有效性。

2.3.5 鼻咽癌靶区勾画数据集

由于本文第四部分研究话题较新, 目前缺少大规模带标注的公开数据集能够用于模型训练和验证, 本课题首先从五家医院回顾性的收集了 1227 个鼻咽癌患者 (Nasopharyngeal Carcinoma, NPC) 的初诊 MRI 数据用于本课题研究。根据实际

① <https://www.creatis.insa-lyon.fr/Challenge/acdc/databases.html>

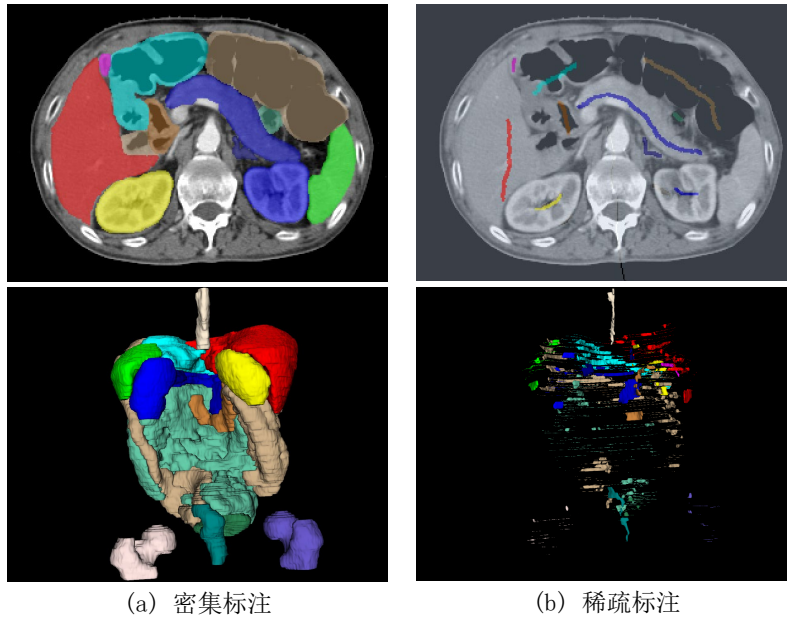


图 2-12 腹部器官分割任务中密集标注和稀疏标注（涂鸦）的三维差异示例，
第一行和第二行分别显示了 2D 和 3D 空间中的可视化

临床需求，本课题邀请了多个高年资肿瘤科医生对患者的原发灶（Primary Gross Tumor Volume, GTV）进行了标注^[176]。其次根据数据分布和标注风格的差异，本课题将这批数据拆分为多中心数据集和多标注者数据集，用于验证深度学习模型在多中心、多标注者鼻咽癌靶区分割任务上的泛化性和鲁棒性，具体细节如下。

1) 多中心鼻咽癌靶区勾画数据集：多中心鼻咽癌靶区勾画数据集主要是包含了来自 5 家医院的 1057 例鼻咽癌病人，主要包括来自南方医科大学附属南方医院的 367 例（SMU）、华西医院的 284 例（WCH）、四川省人民医院的 208 例（SPH）、安徽省省立医院的 146 例（APH）和四川省肿瘤医院的 52 例（SCH1）；其主要临床信息及 MR 成像参数如表 2-2 的多中心数据集部分所示。为保证标注标准的一致性，所有 1057 数据标注均由四川省肿瘤医院的一名资深主治医师手动勾画并由另一位主任医师检查和确认。此步骤能在研究数据分布差异对模型泛化的影响，有效减少标注风格带来的影响。图 2-14 展示了多中心鼻咽癌数据集中不同中心数据的分布特点，体现了该数据集的数据分布存在较大差异。

2) 多标注者鼻咽癌靶区勾画数据集：多标注者鼻咽癌靶区勾画数据集主要是由来自四川省肿瘤医院（SCH2）的 170 例鼻咽癌病人组成，其主要临床信息及 MR 成像参数如表 2-2 中多标注者部分所示。不同于多中心鼻咽癌靶区勾画数据集，本数据集旨在研究本章提出的方法能不能有效缓解不同标注者的标注差异带来的模型性能下降。为此，本文邀请了来自另外医院和医疗组的两位主治医师和副主任医师分别独立的进行鼻咽癌原发灶勾画。表 2-3 展示了这些标注专家的基本信息，

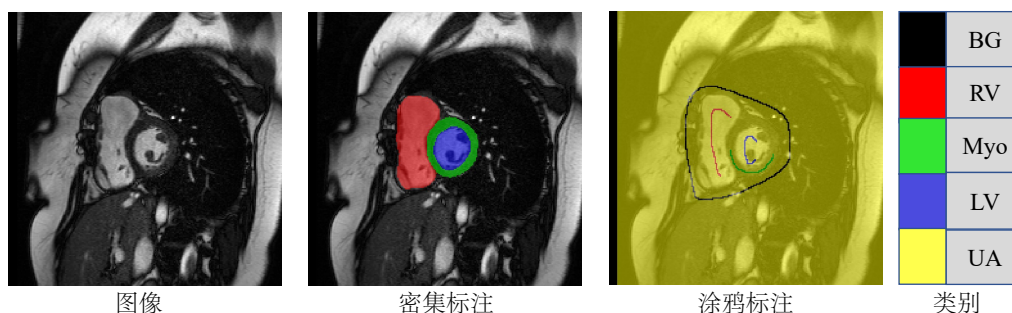


图 2-13 心脏结构分割任务中像素级标注和稀疏标注（涂鸦）的差异示例。

BG、RV、Myo、LV 和 UA 代表背景、右心室、心肌、左心室和未标注像素

他们分别来自 SCH、SMU、SPH 和成都市锦江大观医院（DGH），这些信息充分说明这些医生的临床经验较丰富，勾画结果比较可靠。表2-4展示了这些标注专家之间的标注差异，从表可以看出不同标注者对鼻咽癌原发灶勾画存在显著差异，全部 Dice 均小于 85%，95HD 也大于 7 mm，凸显了该数据集不同标注者之间标注风格存在较大差异。

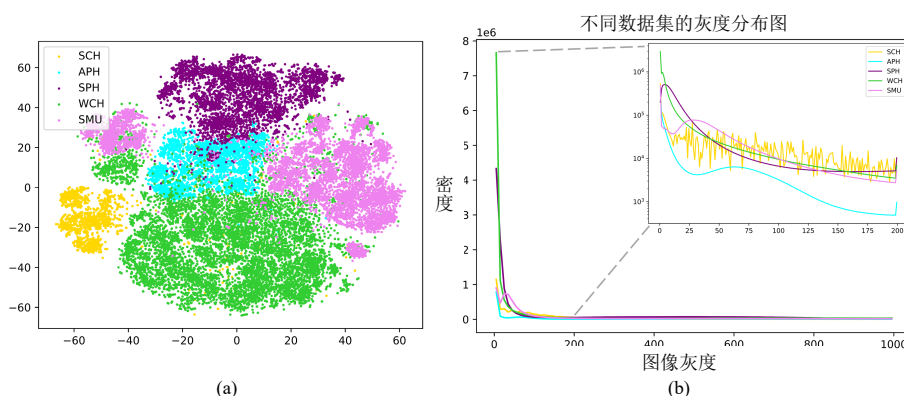


图 2-14 多中心鼻咽癌数据集数据特性图。(a) 不同中心数据的特征分布的 t-SNE^[4] 可视化图；(b) 不同中心数据的平均灰度分布曲线图。不同颜色表示数据来自不同医疗中心

2.4 本章总结

本章节主要对本研究涉及的相关理论基础、方法发展和数据集特点进行了介绍。首先介绍了图像分割的基础知识，包括与本文高度相关的传统水平集和图割方法以及深度学习基础。然后详细回顾了深度学习分割方法发展过程中具有重大影响力的相关工作，并介绍现有方法存在的局限性，引出本课题的主要研究内容有限标注学习。之后介绍了有限标注学习研究的相关发展过程，包括主要研究问题和方法。之后详细介绍了这些问题和方法的基础知识，包括问题定义和数学表

表 2-2 多中心和多标注者鼻咽癌靶区勾画数据集详细临床信息。* 表示数值
结果为中位数和范围

数据集		多中心			多标注者	
变量	SMU (n=367, %)	WCH (n=284, %)	SPH (n=208, %)	APH (n=146, %)	SCH1 (n=52, %)	SCH2 (n=170, %)
性别						
男性	262 (71.4)	206 (72.5)	134 (64.4)	107 (73.3)	37 (71.2)	120 (70.6)
女性	105 (28.6)	78 (27.5)	74 (35.6)	39 (26.7)	15 (28.8)	50 (29.4)
年龄 *						
T 类别	47 (12-80)	49 (15-79)	49 (28-76)	50 (17-79)	46 (26-74)	49 (15-79)
T1	37 (10.1)	32 (11.3)	16 (7.7)	18 (12.3)	11 (21.1)	30 (17.6)
T2	97 (26.4)	57 (20.1)	40 (19.2)	27 (18.5)	8 (15.4)	40 (23.6)
T3	160 (43.6)	129 (45.4)	115 (55.3)	72 (49.3)	28 (53.8)	68 (40.0)
T4	73 (19.9)	66 (23.2)	37 (17.8)	29 (19.9)	5 (9.7)	32 (18.8)
制造商						
GE	334 (91.0)	284 (100.0)	0 (0.0)	113 (77.4)	0 (0.0)	
Siemens	24 (6.5)	0 (0.0)	207 (99.5)	27 (18.5)	52 (100.0)	170 (100.0)
Philips	9 (2.5)	0 (0.0)	1 (0.5)	6 (4.1)	0 (0.0)	
强度						
1.5 T	336 (91.6)	0 (0.0)	10 (4.8)	119 (81.5)	23 (44.2)	0 (0.0)
3.0 T	31 (8.4)	284 (100.0)	198 (95.2)	27 (18.5)	29 (55.8)	170 (100.0)
回声时间 * (ms)	2.66 (2.04-19.0)	9.47 (3.79-10.39)	9.9 (7.8-15.0)	14.9 (10.0-16.0)	11 (2.03-19.0)	9.47 (3.79-10.39)
重复时间 * (ms)	455 (80-2334)	721 (9.54-931)	788 (450-2190)	610 (309-1223)	756 (4.39-860)	721 (9.54-931)
视野 (cm)*	25 (24-30)	25 (21-26)	23 (21-33)	26 (24-32)	24 (20-30)	26 (24-26)
翻转角度 *	80 (15-180)	111 (12-111)	120 (69-150)	90 (90-160)	150 (9-160)	111 (12-111)
厚度 (mm)						
1-3 mm	4 (1.1)	278 (97.9)	3 (1.4)	108 (74.0)	17 (32.7)	170 (100.0)
>3-5 mm	5 (1.4)	4 (1.4)	15 (7.3)	38 (26.0)	35 (67.3)	0 (0.0)
>5-8 mm	358 (97.5)	2 (0.7)	190(91.3)	0 (0.0)	0(0.0)	0 (0.0)
平面间隔 (mm)						
0.3-0.5 mm	329 (89.6)	280 (98.6)	181 (87.0)	106 (72.6)	10 (19.2)	170 (100.0)
> 0.5-0.8 mm	38 (10.4)	4 (1.4)	27(13.0)	40 (27.4)	42 (80.8)	0 (0.0)

表 2-3 多标注者鼻咽癌靶区勾画数据集的相关标注者信息

医生编号	工作年限	来自医院	治疗病人数
专家 1	12	SCH	> 1000
专家 2	11	DGH	≈ 200
专家 3	13	SPH	> 1000
专家 4	11	SMU	> 1000

表 2-4 不同肿瘤科医生在勾画鼻咽癌原发灶时的差异，以 Dice(%) 和 95HD (mm) 为例

95HD (mm) Dice (%)	专家 1	专家 2	专家 3	专家 4
专家 1	-	8.36 ± 5.52	7.12 ± 5.54	10.02 ± 5.16
专家 2	72.79 ± 10.78	-	7.43 ± 4.84	8.41 ± 4.59
专家 3	74.08 ± 11.37	72.26 ± 11.07	-	9.61 ± 5.88
专家 4	70.38 ± 11.06	72.72 ± 8.56	69.80 ± 12.20	-

达，并分析和讨论了其中主要方法的优缺点。最后介绍了本研究主要涉及到的公开数据集，包括三个 MRI 分割数据集和四个腹部多器官分割 CT 数据集和针对本研究收集建立的大规模多中心、多标注者鼻咽癌靶区勾画数据集。综上，本章内容为后续研究提供了详细的理论和实践基础。

第三章 基于测地线距离变换的可泛化交互式分割标注

3.1 引言

传统器官和病灶分割由资深医生手工完成，人工勾画效率低、耗时长、勾画成本极高，有经验的医生完成一个病人的多器官分割也往往需要花费几个小时 [44, 177, 178]。尽管自动分割方法已经研究了多年，但对于具有大量解剖变异和复杂病理情况的情况，它们仍然很难获得一致精确的分割 [76]。这主要是由于医学图像固有的限制，如低对比度、不同的成像和分割协议以及患者之间的变化 [77]。此外，由于受到医学图像数据规模较小、标注获取依赖临床专业知识和医生的临床经验，使得收集和标注大规模数据来开发高性能模型变得昂贵和困难。相比之下，交互式分割方法利用用户的知识和经验，利用极少的标注数据即可获得更准确和稳健的分割模型，使其更加贴合临床需求并被广泛推广使用 [77, 177-179]。

为了更好的贴近实际应用，一个理想的交互式分割方法应该具备以下几个特性：1) 在尽可能少的用户交互和数据标注情况下实现准确的分割结果，从而减轻标注负担；2) 具有高效性，以便用户可以获得实时响应，即使处理三维数据也是如此；3) 对不同的目标具有良好的泛化能力，以便为新的目标或图像模态产生标注。然而，现有的交互式分割方法很少能够同时满足所有这些挑战性的要求。目前基于深度学习的交互式分割方法获得了越来越多的关注 [76, 77]。但是如何有效地编码用户交互并将其与深度学习模型进行高效融合也是一个实际且具有挑战的问题。大多数现有工作通过将用户交互转化为提示图进行编码，例如欧式距离图 [74, 81, 180-182]、高斯热力图 [82, 183] 以及从用户点变换得到的等高线 [184]。然而，这些编码方法没有考虑图像上下文信息。相比之下，测地线距离变换在空间上平滑且对比度敏感，可以用来编码用户交互 [62, 71, 185]。DeepIGeoS [77] 使用测地线距离变换结合特别设计的阈值来处理用户提供的交互。然而，当处理不同目标时，寻找适当的阈值来截断生成的测地线距离图是耗时的。本章假设一种上下文感知且无需参数的编码方法有助于提高分割的准确性和泛化性。

为了解决上述挑战，本章首先提出了一个新的通用框架，用于更智能和准确地交互式分割二维和三维医学图像，可同时实现对先前见过目标 (小规模训练集数据) 的高性能和高效率分割，并实现对一系列先前未见目标分割的高泛化性。本章的方法充分利用了 CNN 的优势，并只需要少量标注数据进行模型训练和少量的点作为用户交互进行测试。本章提出了一种基于指数化测地线距离变换 (Exponentialized Geodesic Distance, EGD) 的新的用户交互编码方式，它是上下文

感知且无需参数，有助于改善 CNN 获得的分割结果。本章还提出了一种信息融合方法，将额外的用户点与初始分割有效地融合，以获得修正的分割。与现有的交互式医学图像分割框架不同^[76,77,80]，本章的方法更高效，因为它只在图像的子区域上工作，并且不需要对修正进行额外的 CNN 训练。此外，本章通过大量先前见过和未见过的目标验证了该框架的有效性。本章的方法在不同类型的图像背景和模态中验证了对五种类型的二维未见目标和四种类型的三维未见目标的优越性，比现有的交互式分割方法更具优越性。

3.2 算法设计

提出的最小交互深度学习分割框架被称为 MIDeepSeg，如图 3-1 所示。它由两个阶段组成：基于 CNN 和用户交互的初始分割阶段（第一阶段）和基于用户交互和信息融合图割的结果修正阶段（第二阶段）。在图像分割和标注过程中，该方法运行用户多次修正模型，直到分割结果完全可接受。总之，在用少量标注数据进行训练后，本方法可用于先前未见目标的分割和标注，而无需微调或重新训练。

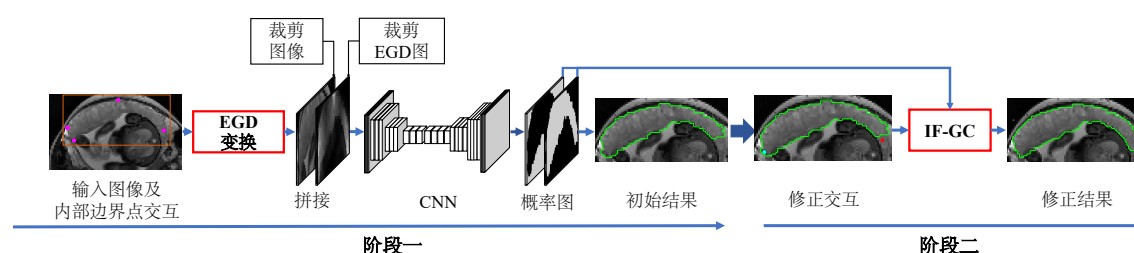


图 3-1 本章提出的最小交互深度学习分割框架（MIDeepSeg）流程图。在第一阶段中（Stage 1）用户提供的内部边界点通过指数化测地线距离（EGD）进行编码，以引导模型获得初始分割结果。在第二阶段中（Stage 2）根据额外的用户修正点和本章提出的信息融合图割（IF-GC）来修正初始分割。请注意，此框架可用于先前未见目标的分割，无需额外的微调或重新训练

3.2.1 基于内部边界点的用户交互

许多现有基于 CNN 的交互式分割框架使用涂鸦^[76]、边界框^[80]或两者兼有^[77]作为交互提示。它们需要用户仔细拖动光标，这需要大量用户交互^[82,186]。正如先前的研究所展示的^[81,82,183,186]，使用点作为用户交互是一种更加用户友好和有效的方式。最近，Maninis 等人^[82]提出了一个框架，只需要用户提供目标的极端点（即，左、右、上、下最边缘的像素）进行 RGB 图像分割，大大减少了用户交互的数量。然而，在医学图像中，获取准确的极端点标注很难且耗时，增加了用户的

负担，因为目标器官或病变在不同患者或成像协议中的大小和形状变异很大，特别是在三维影像数据中。此外，对于不规则和凹形的形状，极端点不足以捕捉目标的主要形状（如图 3-2 所示），这可能会限制 CNN 的性能。为了缓解这些限制，本章提出使用内部边界点作为用户交互，用户只需要提供一些点，这些点位于目标的内侧并且靠近边界。与 DEXTR^[82] 最多使用四个极端点及可选的一个额外点相比，本章的内部边界点可以为具有复杂和不规则形状的不同类型器官提供更多的形状信息。此外，在测试时将点放在目标边界上甚至是极端点上对用户来说很困难，而将点放在边界的内侧使得交互更加友好和方便实现，能够容忍不准确的点。本章将这些点朝着内部区域扩展，因为这些内部边界点的指数化测地线距离变换可以很好地近似分割目标的显著性图，如图 3-3 所示。因此，内部边界点在引导 CNN 处理不同类型未见目标方面具有潜在优势。

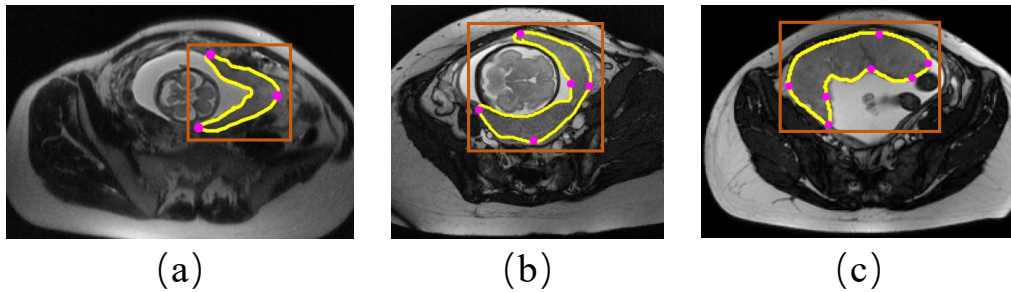


图 3-2 训练数据集上不同形状的胎盘进行内部边界点的模拟。紫红色：在胎盘边缘模拟的点。棕色：内部边界点导出的边界框。黄色：金标准

在训练过程中，所有分割目标的内部边界点都是基于金标准和边缘检测器^[187]自动生成的。这些内部边界点根据两个规则生成：首先，这些点应位于目标内部并靠近边界。其次，由这些点确定的边界框应覆盖整个目标区域。因此，本章在训练图像中模拟用户交互的过程分为两步：1) 为确保边界框覆盖感兴趣区域的整个区域，本章从接近目标目标的极端点^[82]附近选择金标准边界上的少量点（对于二维目标为三或四个，对于三维目标为五或六个）。然后，本章从目标的其余边界点中随机采样 n 个点以提供更多的形状信息，其中 n 是从 0 到 5 的随机数；2) 为了模拟可能不准确地在目标边界上的真实用户点，将步骤 1 中获得的所有这些点稍微向边界内侧扩展几个像素/体素以获得本章的内部边界点。然后，由这些点确定的边界框被向外扩展了几个像素/体素以包含一些背景区域。图 3-2 显示了训练图像上模拟的二维内部边界点和边界框的示例。在测试阶段，要求用户以满足上述两个规则的方式提供内部边界点。用户交互确定的边界框会扩展几个像素以包含一些背景信息。

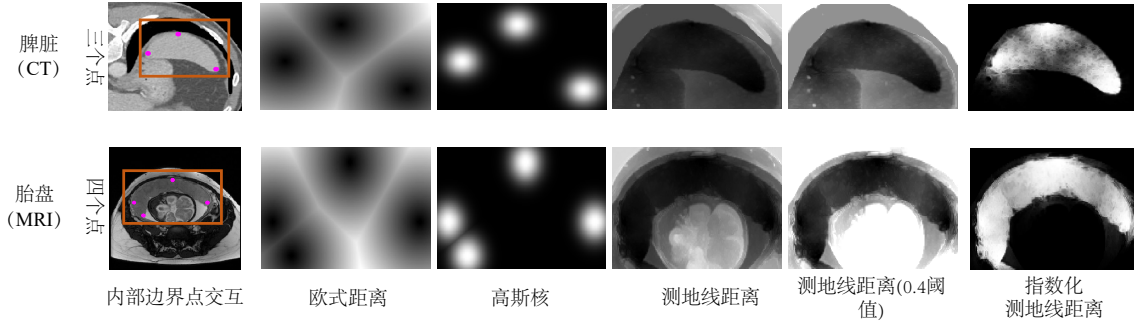


图 3-3 从用户提供的内部边界点生成的不同提示图的可视化比较。(紫红色：内部边界点。棕色：边界框)

3.2.2 指数化测地线距离

对于基于 CNN 的交互式方法来说，高效地编码用户交互至关重要。理想的编码方法应考虑图像上下文，并且可以直接与 CNN 结合，而无需任何手动设计的参数。然而，现有的交互编码方法，如欧式距离变换^[81,180]、高斯热图^[82,183]、等高线^[184]和测地线距离变换^[77]，在同一时间内并不具备这些优点。为了解决这个问题，本章提出了一种上下文感知且无需参数的编码方法：指数化测地线距离（EGD）变换，它是测地线距离变换和指数变换的组合。

假设 $\mathcal{S}_s$ 表示训练阶段模拟的内部边界点或测试阶段用户提供内部边界点所属的像素/体素集合。设 i 是输入图像 I 中的一个像素/体素，则从 i 到 $\mathcal{S}_s$ 的无符号 EGD 为：

$$EGD(i, \mathcal{S}_s, I) = \min_{j \in \mathcal{S}_s} e^{-\mathcal{D}_{geo}(i, j, I)} \quad (3-1)$$

$$\mathcal{D}_{geo}(i, j, I) = \min_{p \in \mathcal{P}_{i, j}} \int_0^1 \|\nabla I(p(n)) \cdot \mathbf{v}(n)\| dn \quad (3-2)$$

其中， $\mathcal{P}_{i, j}$ 是像素/体素 i 和 j 之间的所有路径的集合。 p 是一条可行路径，并且由 $n \in [0, 1]$ 参数化。 $\mathbf{v}(n) = p'(n) / \|p'(n)\|$ 是沿着路径方向的单位向量。请注意，这里定义的 EGD 适用于标量图像，但可以很容易地扩展到矢量值（即多通道或多模态）图像。图 3-3 显示了应用于一些内部边界点的不同编码方法得到的一些提示图的示例。可以观察到，基于 EGD 的提示图比基于其他编码方法的更好地区分前景和背景。因此，它有潜力为 CNN 提供更多的形状、位置和上下文信息，以引导其获得良好的初始分割结果。

3.2.3 基于提示图和模型的初始分割

在本文中，本章专注于设计一个高效且通用的框架来处理不同类型图像中的已见和未见目标。因此，本章的框架不依赖于特定的 CNN 结构设计。为了证明其实用性，本章分别使用了微调后的 2D-UNet^[8] 和 3D-UNet^[32] 进行二维和三维分割。本章将批归一化层替换为实例归一化层，这对不同类型的图像具有更好的适应性，并将特征通道数减少了四分之一，以平衡性能、内存成本和时间消耗。在训练阶段，所有内部边界点和边界框都是基于金标准自动模拟的，如第3.2.1节所述。然后，所有内部边界点都被转换成一个提示图，该提示图与裁剪后的输入图像串联作为 CNN 的输入，如图 3-2所示。在测试阶段，用户被要求为给定的目标提供内部边界点。然后，CNN 可以给出一个初始的分割结果。为了纠正错误的分割，本章使用了一个修正阶段，在初始分割和额外的用户点之间进行信息融合，如下所述。

3.2.4 基于初始分割和修正点信息融合的结果修正

对于基于深度学习的交互式分割，支持修正初始分割结果至关重要。现有方法要么需要一个额外的模型进行修正^[77,84,85,167,168]，要么需要对预训练模型进行特定图像的微调^[76]。然而，这些修正方法耗时、耗内存，且不适用于未见过的目标。此外，Chen 等人^[130] 和 Kamnitsas 等人^[135] 使用 CRF^[188] 来自动修正 CNN 的预测。然而，这些基于 CRF 的^[188] 修正方法^[130,135] 并不是为交互式分割而设计的。与这些方法不同，本章提出了一种基于初始分割和额外用户交互信息融合的高效简单的修正方法，该方法在不需要额外的微调和重新训练的情况下对未见过的目标具有更好的泛化性。图 3-4显示了本章信息融合方法的示意图。在结果修正阶段，为了修正错误分割结果，该方法需要用户提供一些点来区分过分割和欠分割。为了有效地编码这些新的交互，本章再次使用提出的 EGD 变换来获取两个额外的基于交互的提示图： E^f 和 E^b 分别是基于用户提供的前景和背景点用于修正的 EGD 提示图。请注意，本章不直接使用第一阶段获得的初始 EGD 图，而是将初始内部边界点与修正点相结合，用于计算修正步骤中的新 EGD 图。 E^f 和 E^b 的值在 $[0, 1]$ 范围内，可以被视为前景交互点和背景交互点与其他像素之间的差异。另外 P^f 和 P^b 分别表示由 CNN 获得的初始前景和背景预测概率图。进一步，本章提出了一种信息融合策略来根据修正交互编码图 E^f 和 E^b 来分别修正初始模型预测概率图 P^f 和 P^b 。具体而言，本章的目标是设计自动融合算法根据像素 i 与修正点的差异性（ E^f 和 E^b ）来矫正初始模型预测（ P^f 和 P^b ），即靠近交互点的区域收到几乎影响比原

理的区域影响大。本章定义像素 i 为矫正后的前景 (R_i^f) 和背景 (R_i^b) 概率如下:

$$E_i^f = \frac{e^{-D_i^f}}{e^{-D_i^f} + e^{-D_i^b}} \quad (3-3)$$

$$E_i^b = \frac{e^{-D_i^b}}{e^{-D_i^f} + e^{-D_i^b}} \quad (3-4)$$

$$R_i^f = (1 - \alpha_i) * P_i^f + \alpha_i * E_i^f \quad (3-5)$$

$$R_i^b = (1 - \alpha_i) * P_i^b + \alpha_i * E_i^b \quad (3-6)$$

$$\alpha_i = e^{-\min(D_i^f, D_i^b)} \quad (3-7)$$

其中 $\alpha_i \in [0, 1]$ 是自适应权重因子。当 i 接近修正点时, α_i 接近于 1.0, R_i^f (R_i^b) 更受到 E_i^f (E_i^b) 的影响。当用户没有提供前景 (背景) 修正交互时, 则将相应的 D_i^f 或 D_i^b 设为常数。这里用 C^f 和 C^b 分别表示前景和背景交互点, 因此整个修正点集合为 $C = C^f \cup C^b$ 。设 c_i 表示修正点中像素的用户提供的标签, 则当 $i \in C^f$ 时, $c_i = 1$, 当 $i \in C^b$ 时, $c_i = 0$ 。本章将 R^f 和 R^b 整合到 CRF 中以获得修正后的分割结果:

$$E = \left\{ \sum_i \varphi(y_i | I) + \lambda \cdot \sum_{i,j} \psi(y_i, y_j | I) \mid y_i = c_i \text{ if } i \in C \right\} \quad (3-8)$$

其中 φ 和 ψ 分别是一元能量项和二元能量项。 λ 指定了 φ 和 ψ 之间的相对权重。在本文中:

$$\varphi(y_i | I) = -(y_i \log(r_i) + (1 - y_i) \log(1 - r_i)) \quad (3-9)$$

$$\psi(y_i, y_j | I) \propto \exp\left(-\frac{(I_i - I_j)^2}{2\sigma^2}\right) \cdot \frac{1}{dist_{ij}} \quad (3-10)$$

其中 r_i 表示像素 i 在 R^f 中的值, 如果 i 属于前景则 $y_i = 1$, 否则为 0。 I_i 和 I_j 表示图像 I 中像素 i 和 j 的强度。 $dist_{ij}$ 是像素/体素 i 和 j 之间的欧氏距离。 σ 是控制强度差异影响的参数。在本文中, 方程 3-8 中的 CRF^[188] 可以通过最大流-最小割的图割算法来求解^[2]。

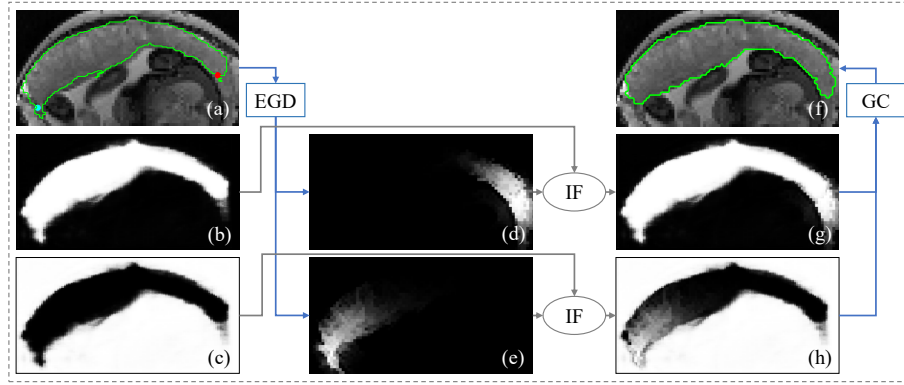


图 3-4 修正信息融合示意图。(a) 用户提供点以指示欠分割（红色）和过分割（青色）区域。(b) 和 (c) 分别是第一阶段 CNN 获得的初始分割前景和背景概率图。(d) 和 (e) 分别是基于前景和背景修正点以及 EGD 转换得到的提示图。(g) 和 (h) 分别是校准的前景和背景概率图。(f) 是修正后的分割结果。其中 IF、EGD 和 GC 分布代表信息融合、指数化测地线距离转和图割

3.2.5 实现细节

本章使用 PyTorch^[189] 分别实现了 U-Net 和 3D U-Net 用于二维和三维图像分割。训练过程在一台配备有 Intel Core i7 CPU 和一块 GTX 1080Ti NVIDIA GPU 以及 120 GB 内存的 Ubuntu(16.04) 上进行。本章使用了 Dice 损失函数和 Adam 优化算法，其中最小批次为 4，权重衰减为 10^{-4} 。对于二维分割，本章总共训练了 300 轮直到网络收敛。学习率在前 150 轮保持为 10^{-4} ，然后每 30 轮减半一次。对于三维分割，本章总共训练了 2000 轮直到网络收敛。学习率在前 1000 轮保持为 10^{-4} ，然后每 200 轮减半一次。每个图像/体积首先基于从内部边缘点推断出的边界框进行裁剪，然后通过裁剪图像的均值和标准差进行标准化。为了提高对未见过目标的泛化能力，在训练阶段本章使用了几种数据增强方法，包括随机旋转、随机缩放、空间和强度的随机翻转。本章使用开源代码计算测地线距离^①和最大流求解方程 3-8^②，并遵循了 DeepIGeoS^[77] 的实验设置。

在本章中，所有带有用户交互的测试过程都在一台配备有 Intel Core i7 CPU 和一块 GTX 1080Ti NVIDIA GPU 的 Ubuntu(16.04) 桌面上进行。按照 DeepIGeoS^[77] 和 BIFSeg^[76] 的做法，方程 3-8 中 λ 的值为 5，方程 3-10 中 σ 的值为 0.1，这是基于验证数据的网格搜索得到的。但是对于特定情况，也允许用户手动设置这两个参数，就像许多现有的方法^[2,24,71] 一样。本章为二维和三维图像开发了两个 PyQt

① 测地线距离: <https://github.com/taigw/GeodisTK>.

② 最大流-最小割: <https://vision.cs.uwaterloo.ca/code/>.

GUI 用于用户交互 3-5^①。

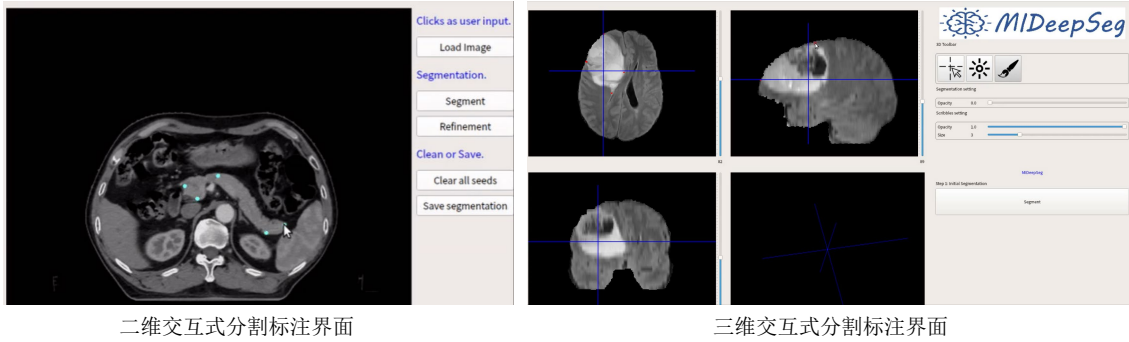


图 3-5 二维、三维交互式分割标注界面

3.2.6 对比方法及评价指标

为了研究不同编码方法在本章分割方法的第一阶段中使用相同的内部边缘点的性能，本章比较了提出的 EGD 与欧式距离变换、高斯变换和测地线距离变换，分别称为 EGD、Eucl、Gauss 和 Geos。此外，本章还将它们与基于无编码交互的边界框分割进行比较，称为 BBox。所有这些方法都基于相同的模型结构。为了公平比较，Eucl、Gauss 和 Geos 均使用了它们各自用于编码用户提供的交互的最佳参数。

MIDeepSeg 与几种现有的交互式分割方法进行了比较。在二维情况下，除了传统方法如 Graph Cuts^[2]、Random Walker^[72] 和 SlicSeg^[64] 之外，本章还比较了最近的基于深度学习的方法，包括 DeepIGeoS^[77]、DIOS^[81]、DeepGrabCut^[190] 和 DEXTR^[82]，其中网络结构均采用了微调后的 U-Net。对于三维分割，本章将 MIDeepSeg 与 ITK-SNAP^[56] 和 3D Graph Cuts^[2] 进行比较，以及三维版本的 DeepIGeoS^[77]、DIOS^[81]、DeepGrabCut^[190] 和 DEXTR^[82]，它们使用了相同的三维分割网络。其中 DeepIGeoS、DIOS、DeepGrabCut 和 DIOS 允许用户多次修正结果。DeepGrabCut 只允许用户在开始时绘制边界框，并且不支持进一步的交互来进行修正。DEXTR 将极端点作为用户交互，并允许用户一次修正结果。Graph Cuts、SlicSeg、Random Walker 和 ITK-SNAP 是传统的交互式分割方法，无需使用带注释的数据集进行训练，并且具有很高的泛化性。相反，DeepIGeoS、DIOS、DeepGrabCut 和 DEXTR 是基于深度学习的方法，需要标记数据进行训练，而 DeepIGeoS 无法处理未知目标。两名用户分别使用这些交互式框架对每个测试图像进行分割，直到结果在视觉上被认为是可接受的，本章报告了两名用户达到的平均结果。分割结果与由经验丰富的放射科医生手动勾画的金标准进行比较。对

^① 算法演示视频: <https://www.youtube.com/watch?v=eq-tqlJnckE>

于定量评估，本章使用了 Dice 相似系数和平均对称表面距离（Average Symmetric Surface Distance, ASSD）：

$$Dice = \frac{2 \cdot |R_p \cap R_g|}{|R_p| + |R_g|} \quad (3-11)$$

其中 R_p 和 R_g 分别表示模型预测的区域和金标准区域。

$$ASSD = \frac{1}{|S_p| + |S_g|} \left(\sum_{i \in S_p} d(i, S_g) + \sum_{i \in S_g} d(i, S_p) \right) \quad (3-12)$$

其中， S_p 和 S_g 分别表示预测结果和金标准的表面点集。 $d(i, S_p)$ 表示点 i 与表面 S_p 之间的最短欧氏距离。为了探究这些方法的分割效率，本章也列出了每个分割任务的用户时间和用户交互点的数量。

3.3 二维医学图像实验验证

3.3.1 二维实验数据

首先，本章使用两个二维分割数据集来验证提出的算法：从胎儿 MRI 和腹部 CT 中分割胎盘和脾脏。具体地，胎盘数据来自于 30 个孕妇在第二孕期的临床 MRI 扫描，采用了轴向视图，像素间距在 $0.7422 \text{ mm} \times 0.7422 \text{ mm}$ 至 $1.582 \text{ mm} \times 1.582 \text{ mm}$ 之间，切片厚度为 $3\text{-}4 \text{ mm}$ 。每个切片都被重新采样为统一间距 $1 \text{ mm} \times 1 \text{ mm}$ 。本章从 18 个图像中选择了 532 个切片进行训练，从 4 个图像中选择了 111 个切片进行验证，从 8 个图像中选择了 176 个切片进行测试。金标准由经验丰富的放射科医生手动勾画。对于脾脏数据，本章从 BTCV 数据集中随机选择了 47 个体积中的 235 个脾脏切片（每个体积 5 个切片）进行训练，并从 TCIA 数据集中选择了 53 个体积中的 159 个脾脏切片（每个体积 3 个切片）进行测试。其次，为了验证本章方法的泛化能力，本章将仅使用在胎盘 MRI 中训练的模型去分割训练集中不存在的多种类型的器官：1) 从 CHAOS 训练集中 CT、T1 和 T2MRI 中的肾脏图像中，分别随机选择 100 个切片用于验证；2) 从 CHAOS 训练集中 CT、T1 和 T2MRI 中的脾脏图像中，分别随机选择 100 个切片进行验证；3) 来自 MSD 的前列腺分割 MRI 数据集，本章从 24 个病人中随机选择了 72 个切片；4) 来自 HC18 数据集的胎儿大脑超声图像，本章随机选择了 60 个切片用于验证。训练集和测试集的信息列在表 3-1 中。为了处理不同大小的不同器官，本章将裁剪的子区域和提示图大小插值到 64×64 作为 CNN 的输入。

表 3-1 用于训练和测试二维交互式分割算法的数据集。注意，对于脾脏，BTCV 和 TCIA 分别是训练集和测试集

分割目标	数据模态	训练集数量	测试集数量	数据集来源
胎盘	MRI (T2)	532	176	伦敦大学学院
脾脏	CT	235 slices	159	BTCV, TCIA
肾脏	MRI (T1)	-	100	CHAOS
肾脏	MRI (T2)	-	100	CHAOS
肾脏	CT	-	100	CHAOS
脾脏	MRI (T1)	-	100	CHAOS
脾脏	MRI (T2)	-	100	CHAOS
脾脏	CT	-	100	CHAOS
前列腺	MRI (T2)	-	72	MSD
胎儿大脑	Ultrasound	-	60	HC18

3.3.2 阈值对不同编码方法的影响

Xu 等人^[181] 和 Li 等人^[180] 使用欧式距离来编码用户交互。距离变换在精心确定的阈值处被截断，以实现有效的表示。然后，截断的提示图与输入图像连接在一起，引导 CNN 获取分割结果。在这些工作之后，本文将所有提示图的数值重新缩放到 $[0, 1]$ 并根据一个合适的阈值将其截断。为了找到合适的阈值，本章使用四个阈值来分别截断欧式距离图分析模型结果，即 0.2、0.4、0.6 和 0.8。使用它们对应的提示图分别引导 CNN 分割 MRI 中的胎盘和 CT 中的脾脏。表 3-2 列出了不同截断欧式距离图用于胎盘和脾脏分割的定量评估结果。实验发现对于胎盘来说最佳阈值为 0.6，而对于脾脏则为 0.4；因此，后续实验中均使用了这些最佳值。Maninis 等人^[82] 和 Wang 等人^[183] 使用高斯变换来处理用户交互。为了生成高斯

表 3-2 使用同一组真实用户交互点，对欧式距离变换的不同阈值进行胎盘和脾脏分割的定量评估

	胎盘 (MRI)		脾脏 (CT)	
	Dice (%)	ASSD (pixel)	Dice (%)	ASSD (pixel)
Eucl (阈值 = 0.2)	86.82±6.65	3.74±2.58	92.88±3.81	2.47±1.58
Eucl (阈值 = 0.4)	86.91±7.20	3.62±2.47	93.58±6.98	2.29±1.13
Eucl (阈值 = 0.6)	87.56±5.98	3.42±2.30	92.62±3.13	2.30±1.16
Eucl (阈值 = 0.8)	87.13±6.26	3.57±2.29	93.04±5.00	2.33±1.08

热图，本文需要为每个任务确定适当的 σ 值。在这项工作中，分别为胎盘和脾脏分割使用了四个不同值的 σ 来生成高斯热图。这些 σ 值分别为 3、6、9 和 12。它们对应的热图与输入图像拼接在一起，引导 CNN 实现初始分割。结果列

在表 3-3 中。可以发现，胎盘的最佳 σ 值为 9，脾脏的最佳 σ 值也为 9；因此在本文实验中采用了这些最佳值。

表 3-3 使用同一组真实用户交互点，对高斯变换的不同 σ 值进行胎盘和脾脏分割的定量评估

	胎盘 (MRI)		脾脏 (CT)	
	Dice (%)	ASSD (pixel)	Dice (%)	ASSD (pixel)
Gauss ($\sigma = 3$)	86.96 $\pm$ 6.89	3.74 $\pm$ 2.58	92.76 $\pm$ 4.16	2.38 $\pm$ 1.52
Gauss ($\sigma = 6$)	87.78 $\pm$ 6.15	3.62 $\pm$ 2.47	92.91 $\pm$ 6.98	2.21 $\pm$ 1.45
Gauss ($\sigma = 9$)	87.91$\pm$6.18	3.42$\pm$2.30	93.22$\pm$3.32	2.29$\pm$1.51
Gauss ($\sigma = 12$)	86.97 $\pm$ 6.79	3.57 $\pm$ 2.29	92.69 $\pm$ 3.70	2.24 $\pm$ 1.26

DeepIGeoS^[77] 使用测地线距离变换对用户交互进行编码。与欧式距离变换一样，本文使用不同的阈值来截断测地线距离图，这些阈值分别为 0.2、0.4、0.6、0.8。截断后的测地线距离图也与输入图像拼接在一起，引导 CNN 实现初始分割。定量评估结果列在表 3-4 中。可以观察到，胎盘的最佳阈值为 0.6，而脾脏的最佳阈值为 0.4。

表 3-4 使用同一组真实用户交互点，对测地线距离变换的不同阈值进行胎盘和脾脏分割的定量评估

	胎盘 (MRI)		脾脏 (CT)	
	Dice (%)	ASSD (pixel)	Dice (%)	ASSD (pixel)
Geos (阈值 = 0.2)	86.81 $\pm$ 6.33	4.00 $\pm$ 2.72	92.96 $\pm$ 3.13	2.23 $\pm$ 0.97
Geos (阈值 = 0.4)	87.08 $\pm$ 5.35	3.59 $\pm$ 2.05	94.02$\pm$3.23	2.13$\pm$0.93
Geos (阈值 = 0.6)	87.17$\pm$6.38	3.55$\pm$2.62	93.06 $\pm$ 4.67	2.17 $\pm$ 0.96
Geos (阈值 = 0.8)	86.93 $\pm$ 7.13	3.62 $\pm$ 2.45	92.91 $\pm$ 3.33	2.23 $\pm$ 0.92

3.3.3 基于提示图和模型的二维初始分割结果

图 3-7 分别展示了使用用户提供的内边缘点进行的 MRI 胎盘和 CT 脾脏的初始分割结果。本章使用相同的用户提供的内边缘点将提出的 EGD 与 BBox、Eucl、Gauss、Geos 进行了比较。请注意，Eucl、Gauss 和 Geos 的参数分别进行了优化以进行比较（优化结果如 3.3.2 节所示）。可以观察到，EGD 变换可以比其他编码方法更准确地引导 CNN 获得分割结果。表 3-5 列出了胎盘和脾脏的不同编码方法的定量评估结果。可以看到，本章的上下文感知和无参数编码方法 EGD 始终优于其他编码方法。二维图像的 EGD 变换的计算时间不到 0.05 秒，可以实现实时响应。图 3-6 展示了在复杂形状的图像中不同数量的内边缘点对初始分割的影响。

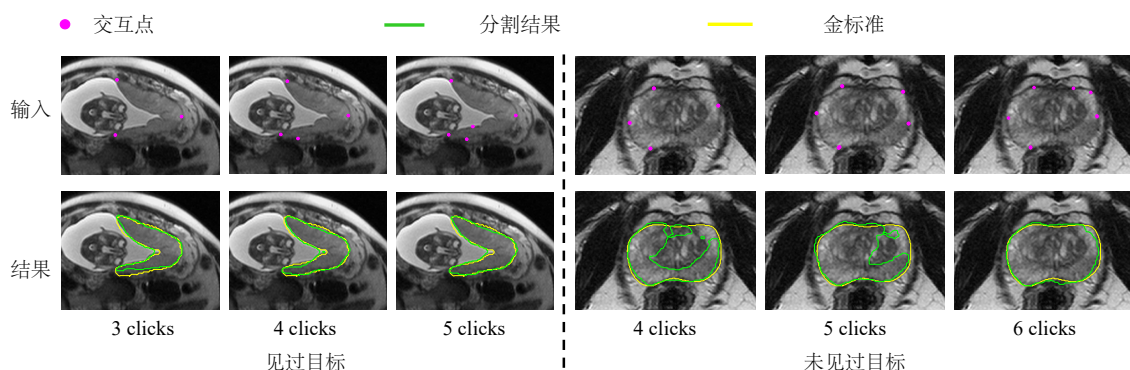


图 3-6 不同数量的初始内边缘点对具有复杂形状的胎盘和前列腺（未见过目标）分割的影响。第一行展示了不同数量内边缘点的输入图像，第二行展示了分割结果

表 3-5 使用同一组内边缘点进行胎盘和脾脏分割时不同编码方法的定量比较。* 表示与第二名方法相比的 p -值 < 0.05

	胎盘 (MRI)		脾脏 (CT)		时间 (s)
	Dice (%)	ASSD (pixel)	Dice (%)	ASSD (pixel)	
Bbox	85.53±7.16	4.37±3.05	91.36±4.69	3.76±1.71	-
Eucl	87.56±5.98	3.42±2.30	93.58±6.98	2.29±1.13	0.001
Gauss	87.91±6.18	3.56±2.43	93.22±3.32	2.29±1.52	0.001
Geos	87.17±6.38	3.62± 1.01	94.02±7.23	2.13±0.94	0.003
EGD	88.10±4.47*	3.33±2.19	95.08±3.23*	2.25±1.28	0.004

3.3.4 基于初始分割和修正点信息融合的二维结果修正

图 3-8展示了使用不同修正方法对胎盘和脾脏分割进行修正的示例。第一行显示了本章框架第一阶段获得的初始分割结果。基于初始分割，本章进一步使用额外的点来获得修正结果。本章比较了使用相同的用户点进行修正的原始图割（GC）和信息融合图割（IF-GC）的修正结果。根据用户提供的点进行了图割，其中“基于初始分割”的分割概率图 and 用户交互（背景和前景种子）作为输入，并通过最大流求解。表 3-6 中列出了胎盘和脾脏分割的性能，其中前两行表明本章在第一阶段的方法已经比相同网络结构的自动分割大大提高了准确性。最后两行表明，相同的用户点情况下本章的 IF-GC 在第二阶段的修正中比 GC 获得了更高的准确性。

本章进一步研究了使用 MIDeepSeg 进行不同目标分割的修正点数量，并在图 3-9中绘制了修正点数量的直方图。本章发现，大部分测试案例不需要额外的交互点即可获得准确结果，只有少数具有挑战性的案例需要进行 4 次以上的修正。

表 3-6 使用相同修正点进行胎盘和脾脏分割结果修正的定量比较。* 表示 IF-GC 与 GC 结果存在显著差异 (p 值 <0.05)

	胎盘 (MRI)		脾脏 (CT)	
	Dice (%)	ASSD (pixel)	Dice (%)	ASSD (pixel)
自动分割结果	79.76 $\pm$ 15.33	8.94 $\pm$ 11.2	90.09 $\pm$ 10.2	10.50 $\pm$ 13.9
一阶段结果	88.10 $\pm$ 4.47	3.33 $\pm$ 2.19	95.08 $\pm$ 3.23	2.25 $\pm$ 1.28
GC 修正结果	88.41 $\pm$ 5.33	3.14 $\pm$ 2.39	95.46 $\pm$ 3.19	2.06 $\pm$ 1.16
IF-GC 修正结果	89.21$\pm$4.37*	2.87$\pm$1.89*	95.79$\pm$3.07*	1.84$\pm$0.86

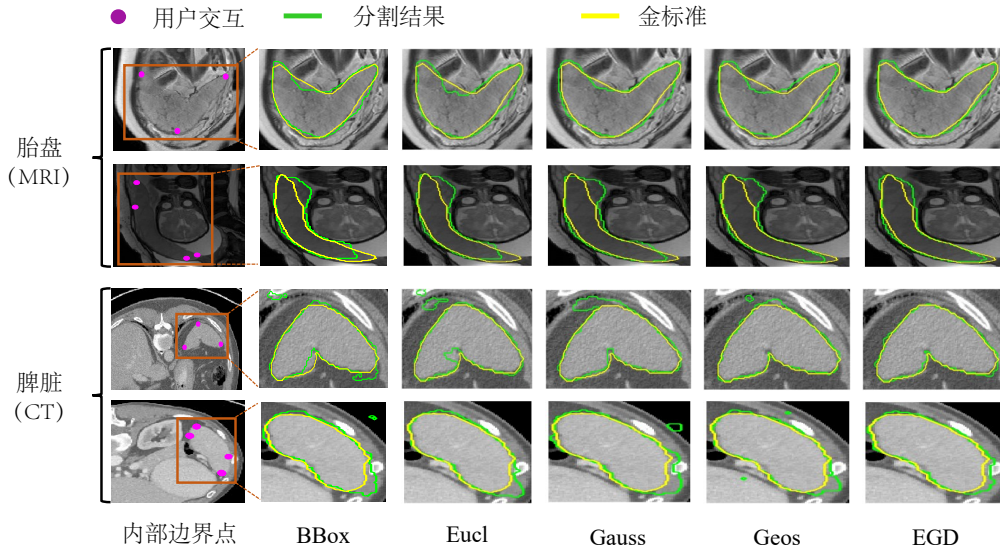


图 3-7 不同编码方法在胎盘和脾脏分割的可视化比较，这是本章方法的第一阶段。第一列显示了带有用户提供的内边缘点（品红色）的图像，其他列显示了初始结果

3.3.5 与其他二维交互式分割方法相比较

本章将 MIDeepSeg 与 DeepIGeoS^[77]、Graph Cuts^[2]、Random Walks^[72]、Slic-Seg^[64]、DIOS^[81]、DeepGrabCut^[190] 和 DEXTR^[82] 分别在胎盘和脾脏分割任务上进行了详细对比。图 3-10 显示了这些方法在二维胎盘分割方面的可视化比较。第一行显示了初始交互和初始分割结果，第二行显示了经过修正后的最终结果和所有用户交互。可以看出，MIDeepSeg 只需较少的用户点就能获得良好的结果，而其他方法需要更多的交互。基于胎盘和脾脏分割结果的定量比较如表 3-7 所示。表明 MIDeepSeg 的准确性高于其他交互式分割方法（除 DeepGrabCut 外），且需要的用户时间和交互点数量都较少。请注意，DeepGrabCut 不允许额外的用户交互进行修正，这导致了它的准确性低于其他方法。这证明了本章提出的方法非常有效地

实现了高度准确的分割结果。

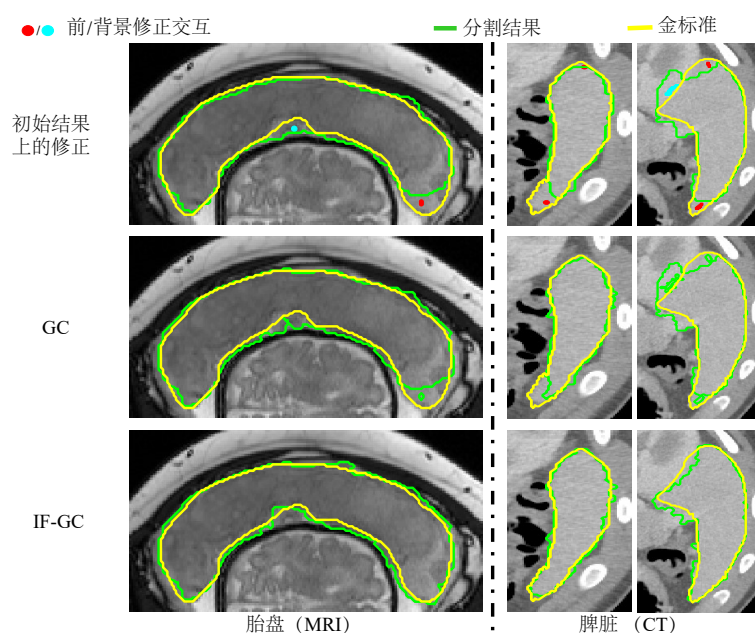


图 3-8 对 GC 和 IF-GC 进行可视化比较。第一行显示了用于修正初始分割结果的用户点。其他行分别显示了通过 GC 和 IF-GC 修正后的结果。这些结果是基于相同的用户点进行修正的。GC 和 IF-GC 分别表示原始图割和信息融合图割

表 3-7 不同交互式方法在 Dice 系数、平均表面距离 (ASSD)、用户时间和交互点数量方面对二维胎盘和脾脏分割的定量比较。* 表示与第二名方法相比的 p -值 <0.05

	胎盘 (MRI)					脾脏 (CT)			
	Dice (%)	ASSD (pixel)	时间 (s)	交互点数量	Dice (%)	ASSD (pixel)	时间 (s)	交互点数量	
Graph Cuts	87.02±5.20	3.12 ±0.42	30.1±10.9	265.0±103.6	95.27±4.36	1.30 ±0.42	21.2±7.7	335.1±91.7	
Random Walks	87.02±4.58	2.95 ±2.66	33.9±34.6	374.3±114.2	95.51 ±1.59	1.45±2.66	20.1±7.9	218.4±69.0	
SlicSeg	87.63±5.71	3.00 ±0.39	25.8±11.5	189.3±81.2	95.18±4.70	1.23±0.39	20.1±8.2	254±77.5	
DeepIGeoS	87.96±5.16	3.89 ±2.74	12.0±8.0	90.6 ± 95.2	96.39±2.22	1.71±2.74	6.1±4.8	31.1±52.4	
DeepGrabCut	86.74±7.03	4.18±2.89	4.2±2.8	2.0±0	92.54±3.36	2.43±1.56	3.8±1.5	2.0±0	
DIOS	87.48±6.31	4.03±2.52	15.3±13.0	12.4±5.7	94.85±2.79	2.06±1.47	7.6±2.7	7.8±4.7	
DEXTR	88.77±4.83	3.07 ±2.25	8.9±3.7	7.2±3.3	94.18±3.25	2.67±1.36	5.9±3.9	5.6±2.7	
MIDeepSeg	89.63±4.15 *	2.69±1.75 *	6.40±3.1	5.75±2.1	96.93±1.43 *	1.18±0.44 *	4.76±2.0	4.85±1.6	

3.3.6 分割未见过的二维医学图像

为了研究 MIDeepSeg 在未见目标上的分割性能和泛化能力,本章将 MIDeepSeg 与现有具有良好泛化能力的方法进行了比较: Graph Cuts^[2]、Random Walks^[72]、SlicSeg^[64]、DIOS^[81]、DeepGrabCut^[190] 和 DEXTR^[82]。对于 MIDeepSeg、DIOS、DeepGrabCut 和 DEXTR, 本章使用仅针对胎盘图像 (T2) 进行训练的模型来从不

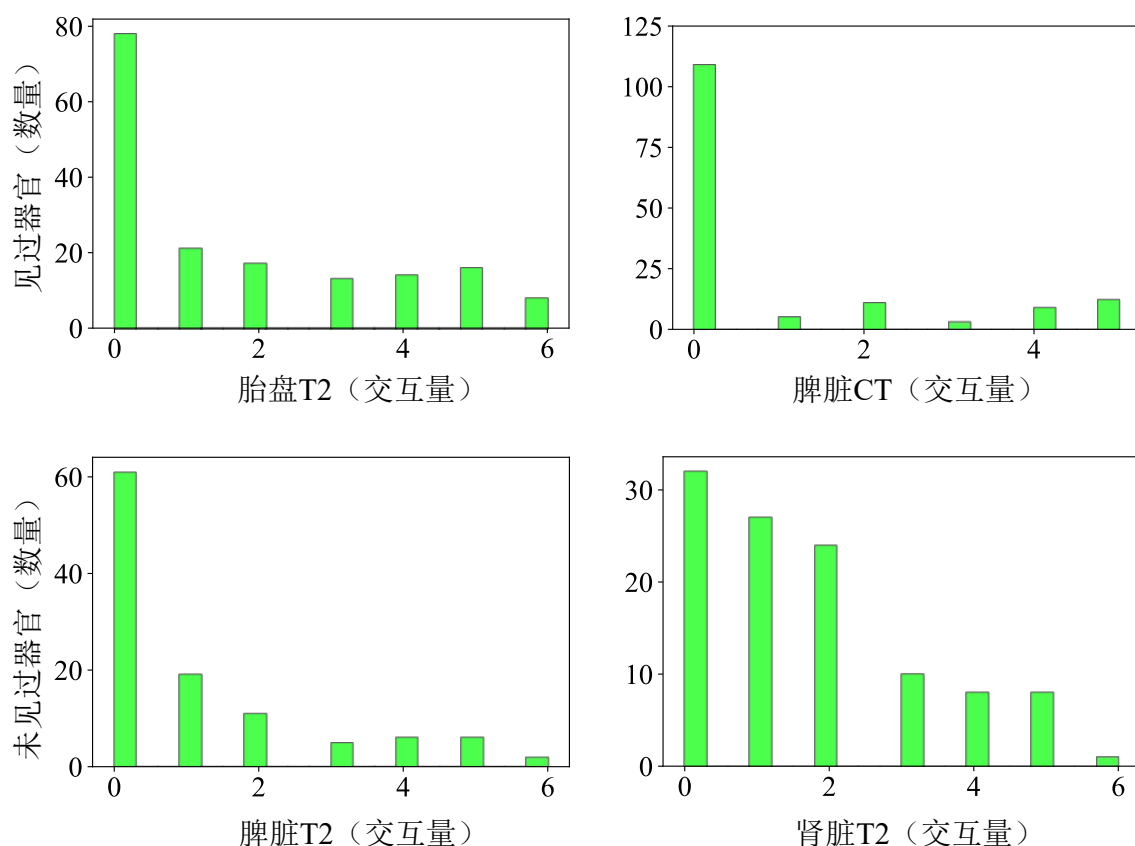


图 3-9 MIDeepSeg 对不同目标所需的修正点数量直方图。在第一行中，胎盘 T2 和脾脏 CT 是见过的目标，而在第二行中，脾脏 T2 和肾脏 T2 是以前未见过的目标

同模态图像中分割四种以前未见的器官（即肾脏、脾脏、前列腺和胎儿脑），量化结果如表 3-1 所示。图 3-11 展示了 MIDeepSeg 对以前未见目标的分割结果。第一行展示了初始交互和初始分割结果，第二行展示了所有交互和最终分割结果。可以观察到，MIDeepSeg 只需少量用户点即可在未见过的器官上获得良好的结果。最终分割结果的定量比较如图 3-12 所示。结果显示，与其他交互式分割方法相比，MIDeepSeg 所需的用户时间和交互明显较少，并且具有相似或更高的准确性。此外，可以观察到，MIDeepSeg 可以很好地处理不同类型的以前未见的图像模态和器官，而无需任何额外的训练或微调。本章进一步研究了使用 MIDeepSeg 进行肾脏（T2）和脾脏（T2）分割时的修正交互数量，并在图 3-9 中展示了交互量的直方图分布。可以发现尽管这些目标不在训练集中，本文提出方法不需要或少量交互即可获得准确的结果。

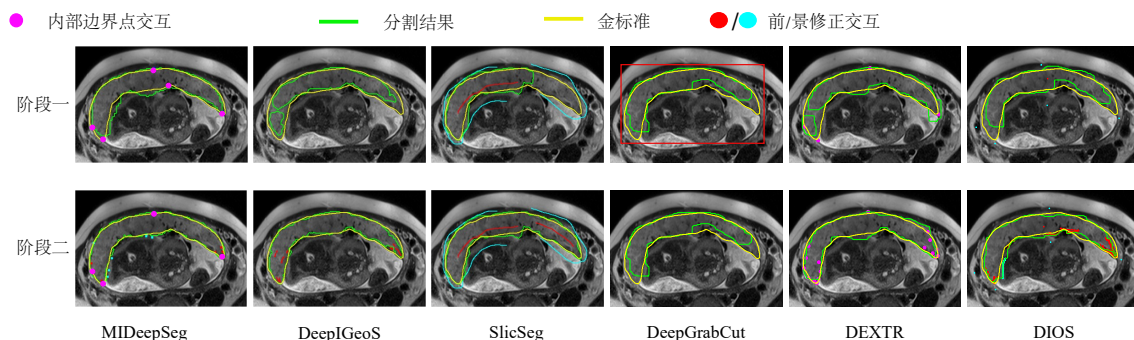


图 3-10 MIDeepSeg 与其他交互式方法在胎盘分割方面的可视化比较。第一行展示了带有或不带有初始交互的初始分割结果。第二行显示了修正后的最终结果

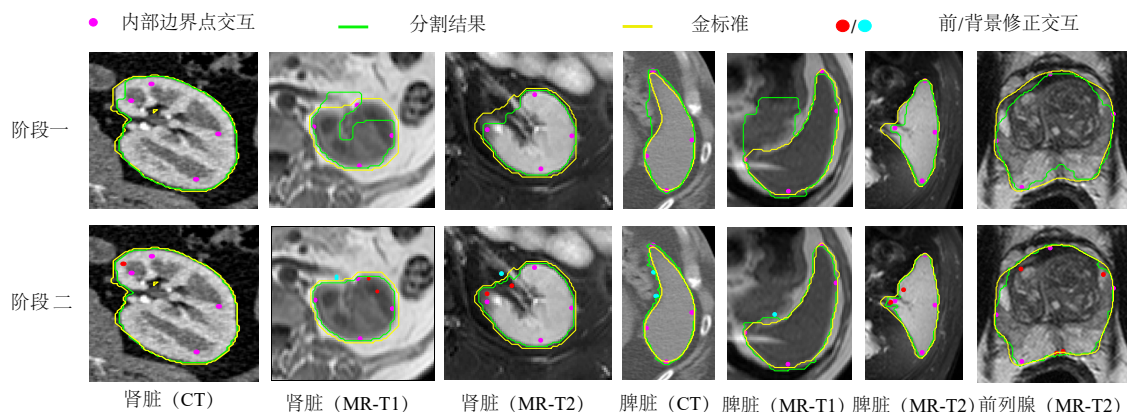


图 3-11 MIDeepSeg 对二维未见过器官分割结果的一些示例。第一行显示了初始用户交互和初始分割。第二行显示了所有用户交互和最终的分割结果。请注意，该模型仅使用 T2 MRI 中的胎盘进行训练

3.4 三维医学图像实验验证

3.4.1 三维实验数据集

首先，本章调研了 MIDeepSeg 从 T1C 图像中分割肿瘤核心的表现。本章使用了 BraTS2018^①训练集，该训练集包含了 285 个病例，具有四种模态：FLAIR、T1C、T1 和 T2。所有图像都已经去除了颅骨并重新采样到 $1\text{ mm} \times 1\text{ mm} \times 1\text{ mm}$ 的各向异分辨率。本章使用 170 个 T1C 图像进行训练，以及 47 个 T1C 图像进行测试。然后，本章进一步探索了 MIDeepSeg 在分割三个不同模态中未出见目标的泛化能力：1) 来自 BraTS2018 的 FLAIR 图像中的整个脑肿瘤，本章随机选择了 60 个病例进

① <https://www.med.upenn.edu/sbia/brats2018.html>

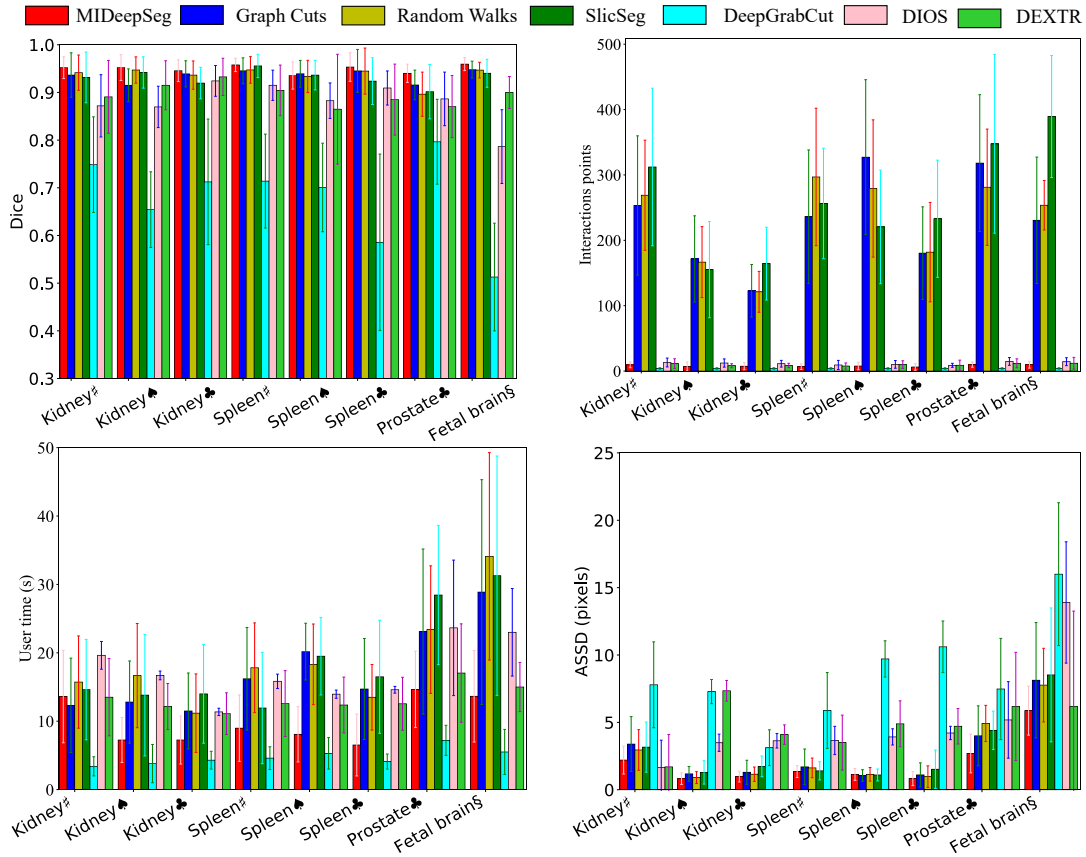


图 3-12 不同交互式分割方法针对未见过的目标的 Dice、ASSD、用户时间和交互数量的比较。其中，♯、§、♠、♣ 分别表示 CT 图像、超声图像、T1 和 T2 图像。所有这些器官在训练集中都未出现过

行测试。2) 来自 KiTS2019^①数据集从 CT 图像中分割肾脏，在这里本章随机选择了 15 个病人（包括 30 个有或没有肿瘤的肾脏）进行测试。3) 来自 ACDC^②的 MRI 图像中的左心室，随机选择了 30 个病例进行测试。KiTS 和 ACDC 的测试数据被重新采样到 $1\text{ mm} \times 1\text{ mm} \times 1\text{ mm}$ 的各向异分辨率。训练和测试数据集的详细分布如表 3-8 所示。为了处理不同大小的三维目标，本实验中将裁剪后的补丁区域和提示图的分辨率调整为 $64 \times 96 \times 96$ 。

3.4.2 基于提示图和模型的三维初始分割结果

为了验证本章提出的 EGD 转换在编码三维内部边缘点的有效性，本章将其与 BBox、Eucl、Gauss 和 Geos 进行了比较。在这个阶段，由用户提供相同的内部边缘点被用于这些方法训练和验证。图 3-13 展示了由不同编码方法引导 CNN 得到的初始分割结果。结果显示，相比其他编码方法，EGD 转换可以更明显地引导 CNN

① <https://kits19.grand-challenge.org>

② <https://acdc.creatis.insa-lyon.fr/>

表 3-8 用于训练和测试的三维实验数据集

分割目标	数据模态	训练集数量	测试集数量	数据集来源
肿瘤核心	MRI (T1C)	170	47	BraTS2018
脑肿瘤	MRI (FLAIR)	No	60	BraTS2018
肾脏	CT	No	30	KiTS
左心室	MRI (T2)	No	30	ACDC

获得更好的分割结果。表 3-9列出了不同编码方法对来自 T1C 图像的肿瘤核分割的定量评价结果。可以观察到，本章的上下文感知和无参数编码方法 EGD 在 Dice 方面始终优于其他方法，为 87.00%，而在 ASSD 方面为 1.46 mm。尽管 EGD 比其他方法需要更多时间（0.24 秒）进行交互编码，但在实践中仍然非常高效。

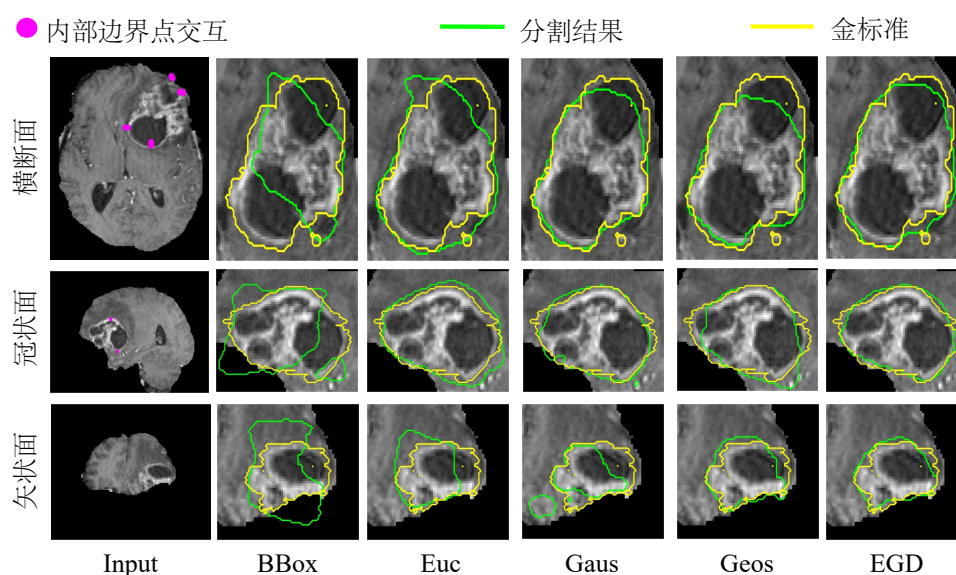


图 3-13 第一阶段不同编码方法分割肿瘤核心结果的可视化比较。所有这些方法都使用相同的内部边缘点和边界框作为输入

3.4.3 基于初始分割和修正点信息融合的三维结果修正

根据本问方法得到的上述初始分割，本研究进一步利用额外的点来获得修正后的结果。本章将简单的图割（GC）和本文提出的信息融合图割（IF-GC）在相同的用户交互式点上进行了比较。肿瘤核心分割的性能如表 3-10所示，显示出本章提出的信息融合方法比其他变体有更高准确性。图 3-14展示了不同修正方法对肿瘤核心分割的一个示例。可以观察到，相同的修正点下，IF-GC 比 GC 更准确地修正了初始分割结果。

表 3-9 在相同的内部边缘点和边界框情况下, 使用不同编码方法进行肿瘤核心分割的定量比较。结果基于 MIDeepSeg 的初始分割 (第 1 阶段)。* 表示与第二名方法相比 p -值 < 0.05

	Dice (%)	ASSD (mm)	时间 (s)
BBox	82.32±12.03	2.17±1.53	-
Eucl	85.25±9.78	1.71±1.20	0.05
Gauss	85.90±9.11	1.64±1.19	0.06
Geos	86.42±8.91	1.60±1.15	0.15
EGD	87.00±9.11*	1.46±1.14*	0.24

表 3-10 不同修正方法在相同的修正点下对三维肿瘤核心分割的定量比较。修正前的分割是在第一阶段由 MIDeepSeg 获得的。GC: 3D Graph Cuts; IF-GC: 信息融合图割; * 表示性能显著高于 GC, p 值 < 0.05

	Dice (%)	ASSD (mm)
自动分割结果	78.08±13.56	2.78±2.22
一阶段结果	87.00±9.11	1.46±1.14
GC 修正结果	87.44±8.31	1.37±1.15
IF-GC 修正结果	88.21±7.31*	1.28±0.94

3.4.4 与其他三维交互式分割方法相比较

图 3-15展示了 MIDeepSeg、3D Graph Cuts^[2]、ITK-SNAP^[56]、以及三维版本的 DeepIGeoS^[77]、DIOS^[81]、DeepGrabCut^[190] 和 DEXTR^[82] 之间的可视化比较。可以发现, MIDeepSeg 只需要少量的内部边界点作为初始交互, 但其初始分割更准确, 并且需要更少的用户交互来获得准确的最终结果。这些方法取得结果的定量比较如表 3-11所示。可以发现 MIDeepSeg 的准确性显著高于其他方法。此外, MIDeepSeg 对于肿瘤核心分割的整个过程平均只需要 29 秒, 是除 DeepGrabCut 外最快的交互式分割方法。

3.4.5 分割未见过的三维医学图像

为了研究 MIDeepSeg 在处理以前未见的三维目标上的泛化能力, 本研究使用了从 T1C 图像中训练的三维 CNN 模型来处理三个以前未见的对象和模态: FLAIR 中的整个肿瘤; CT 中的肾脏和 MRI 中的左心室, 结果如表 3-8所示。两名用户分布

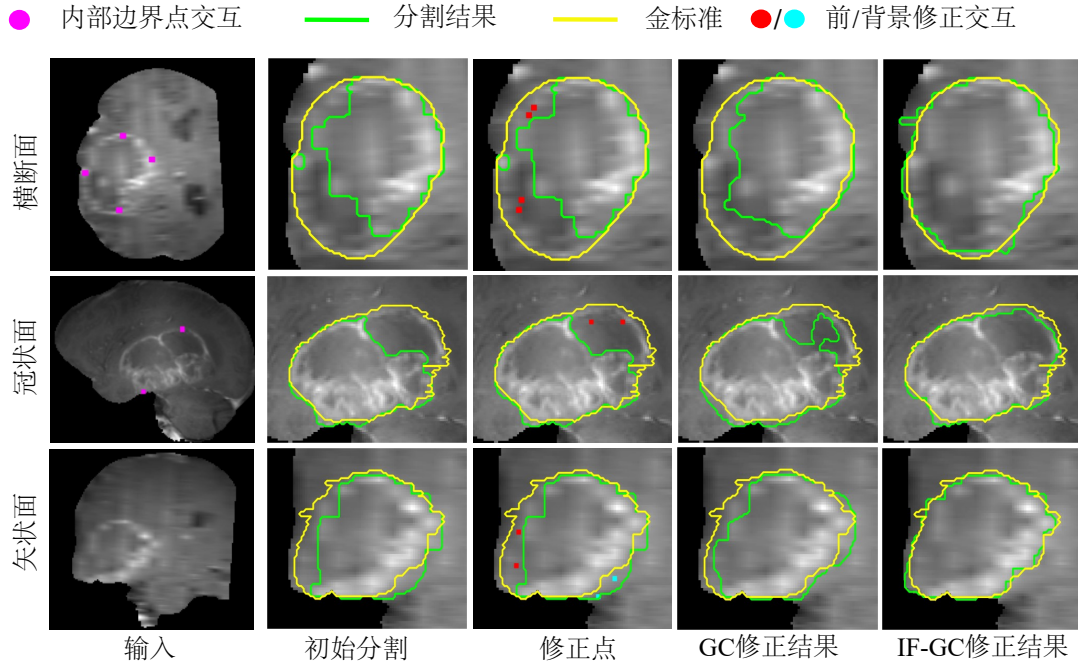


图 3-14 不同修正方法对三维肿瘤核心分割的可视化比较。这些修正方法在相同的初始分割和相同的修正点条件下进行了比较

使用 MIDeepSeg 和两种现有高泛化能力的方法，包括 ITK-SNAP^[56] 以及 3D Graph Cuts^[2]、DIOS^[81]、DeepGrabCut^[190] 和 DEXTR^[82] 来分割这些对象。图 3-16展示了使用 MIDeepSeg 进行三维整个肿瘤、肾脏和左心室分割的一些示例。可以发现，通过少量点，就可以获得不同类型的未见对象的准确结果。定量评估结果如图 3-17所示。它显示 MIDeepSeg 与 3D Graph Cuts、ITK-SNAP、DeepIGeoS、DIOS、DeepGrabCut 和 DEXTR 相比实现了类似或更高的准确性。然而，MIDeepSeg 花费的用户时间显著较少。

3.5 讨论与分析

尽管一些基于深度学习的交互式分割的最近工作^[77,167,168,191]已经表现出良好的性能，但对于当前的 CNN 来说，很难很好地推广到以前未见过的对象类别，因为它们依赖于标注图像来直接学习^[178]。对于医学图像来说，标注图像非常宝贵且稀缺，因为准确的注释需要专业知识和时间。这限制了 CNN 处理未见过的对象的性能，因为这些对象不在训练集中。与传统的 CNN^[8,32] 和迁移学习^[76,192] 相比，本章提出的框架的主要优势是它可以在不重新训练或微调的情况下分割未见过的对象。因此，它显著减轻了收集和标注数据的负担，并且可以直接应用于分割或注释未见过的对象。与 DeepIGeoS^[77] 和 BIFSeg^[76] 相比，MIDeepSeg 只需要少量

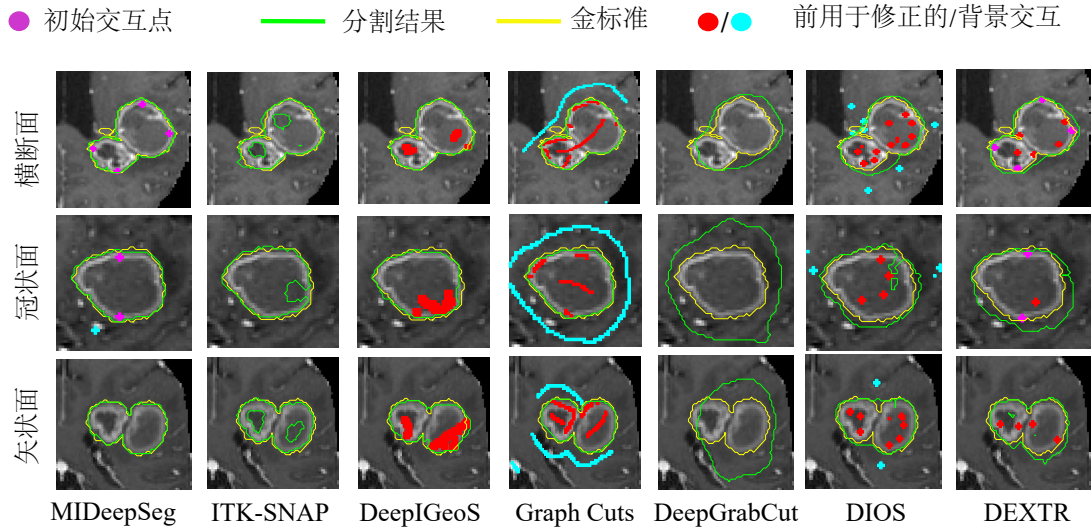


图 3-15 MIDeepSeg、DeepIGeoS、3D Graph Cuts 和 ITK-SNAP 对肿瘤核心分割结果的可视化比较

表 3-11 不同交互式分割方法在肿瘤核心分割任务上 Dice 和 ASSD 的定量评估。* 表示与第二名方法比较时的 p 值 <0.05

	Dice (%)	ASSD (mm)	时间 (s)
Graph Cuts	78.91 ± 14.98	3.46 ± 5.10	99.4 ± 36.7
ITK-SNAP	82.34 ± 11.42	1.99 ± 1.31	173.0 ± 75.5
DeepIGeoS	82.47 ± 10.78	3.34 ± 3.81	82.2 ± 44.7
DeepGrabCut	82.58 ± 11.79	2.89 ± 2.37	10.5 ± 8.3
DIOS	83.34 ± 10.47	2.57 ± 1.79	67.5 ± 23.6
DEXTR	86.39 ± 9.03	1.59 ± 1.11	34.7 ± 18.6
MIDeepSeg	$88.71 \pm 7.00^*$	$1.24 \pm 0.88^*$	28.6 ± 12.2

点作为输入，并具有更高的泛化能力。

现有深度学习框架的一个重大挑战是它们几乎无法推广到以前未见过的对象，并且需要额外的重新训练或微调才能对新目标进行分割。BIFSeg^[76] 使用图像特定的微调来提高 CNN 的泛化能力，但它需要对每个测试图像进行微调，这是一个耗时和内存消耗的过程。基于本文提出的内部边界点、EGD 变换和 IF-GC，MIDeepSeg 可以处理不同类型的未见医学图像，而无需额外的微调或训练。

尽管本文的 EGD 实现简单，但它以前尚未被提出用于用户交互编码，并且与测地线距离有两个重要的区别：首先，EGD 是无参数的，具有更高的泛化能力。测地线距离方法^[77] 需要用户定义的阈值来确保交互只影响局部区域，这降低了其泛化能力，因为不同的图像可能需要不同的阈值。相反，本文提出的 EGD 不需要

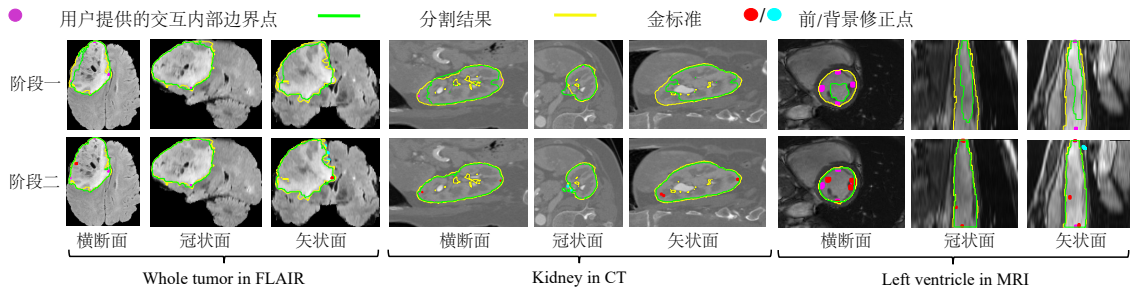


图 3-16 使用 MIDeepSeg 分割未知对象的三个示例。请注意，训练集仅包含 T1C 图像中的肿瘤核心数据

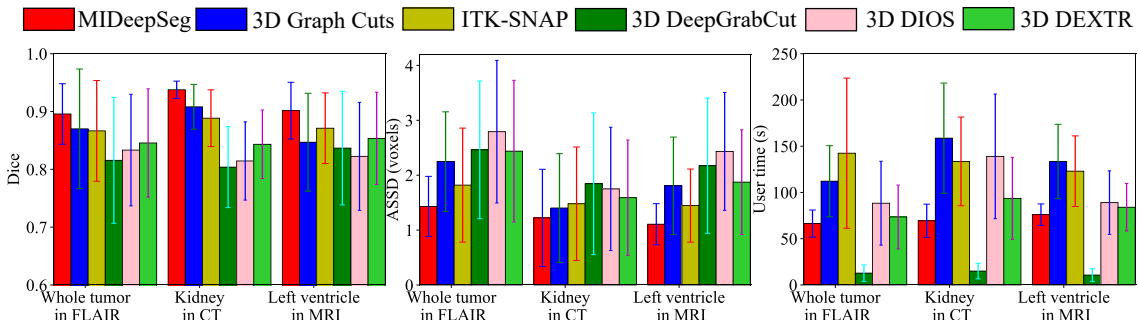


图 3-17 不同交互式分割方法对于三维未见目标分割的 Dice 和 ASSD 及交互时间

这样的参数，可以在不进行特定调整的情况下应用于不同的图像，使其成为一种简单有效的方法，具有更广泛的适用性。其次，EGD 自然地输出概率图，可以用作用户交互指示的前景或背景像素属于的概率。这种概率视角使其能够无缝地集成到条件随机场形式中进行修正。

表 3-5 和表 3-9 展示了 EGD 的计算时间开销，对于二维和三维图像分别需要不到 0.05 秒和 0.25 秒，这对于用户交互的快速响应是可以接受的。本文还研究了本方法在其他阶段的计算时间：二维和三维网络的推理时间分别为 0.008 秒和 0.04 秒。CRF 优化时间分别为 0.015 秒和 0.5 秒。整个用户时间在二维任务情况上约为 8-12 秒（图 3-12），在三维任务下约为 60-80 秒（图 3-17）。因此，本章的方法对于未见对象的交互式分割是高效的。

本文实验发现基于校准概率图和图割的修正方法在不同器官和不同模态的各种情况下表现良好。其优点包括：1) 修正步骤与基于 CNN 的初始分割步骤不存在强相互依赖，因此可以作为一个通用的修正工具，用于交互式地纠正不同网络得到的分割结果以及未见对象。2) 它计算效率高，允许实时响应用户交互，这对于改善交互式分割的用户体验非常重要。3) 用户交互被用作硬约束，确保用户给定的点在修正后将具有其期望的标签。一个潜在的问题是，在复杂情况下可能需

要相对较多的点才能获得准确的结果。然而，在实践中，本章的方法易于使用，并且在处理不同未见对象时效率高，正如实验结果所示。

交互式分割的一个普遍问题是结果可能取决于用户的知识和经验，因为用户会不断修正分割，直到视觉上可接受为止，这标准可能是主观的。然而，本文的方法对用户交互有一些要求：在第一阶段，交互需要在接近内侧边界的位置给出，在第二阶段，交互只在不正确的区域给出，对于大多数情况来说，不正确的区域很小，导致不同用户提供的点范围受到限制，因此用户间变化很小。由于本方法不要求用户精确地在边界或极端点上提供点，且内部边界点可容忍不准确的点，这更加用户友好。正如图 3-11 中所示，第一列的内部边界点不准确且远离边界，在第四列中，顶部点也不准确，甚至在背景中被点，但它们仍然导致良好的初始分割结果。这进一步证明了 MIDeepSeg 的鲁棒性和泛化能力。最近，一些工作^[193-196]使用 Fisher 信息、自然语言、主动学习和深度强化学习来开发智能交互式分割或标注工具。在未来，利用主动学习^[197]和深度强化学习^[167]以及不确定性估计^[191]来引导用户交互进行修正是很有意义的，这有潜力进一步提高交互式分割的效率。

3.6 本章总结

本章提出了一个基于深度学习的交互式分割标注框架，具有对未见对象的良好泛化能力，用于医学图像分割，仅需要少量点作为用户输入。首先提出了一种新颖的上下文感知且无参数的编码方法，用于编码用户交互以引导 CNN 进行良好的初始分割。基于编码方法，其次还提出了一种有效的修正方法，用于提高分割结果的准确性。该框架旨在提高对未见对象的泛化能力，这对于基于深度学习的模型来说非常重要。对于从各种二维和三维图像中分割广泛的以前见过和未见过的器官或病变的实验表明：1) 基于内部边界点和 EGD 变换的框架在准确性和效率方面优于现有的基于深度学习的交互式分割工具。2) 所提出的框架在以前未见过的对象上泛化能力良好。它可以用作一个跨模态、跨目标结构的标注工具，以更高的准确性和更高效率的获取一系列对象的分割结果，从而大大减少大量医学图像分割任务中的数据标注成本。

第四章 基于动态混合伪标签的涂鸦监督医学图像分割

4.1 引言

上一章主要调研了基于点和少量标注数据的医学图像交互式分割和标注方法，实验结果表明利用极少的用户交互模型在有限的标注成本下准确和高效地处理来自不同中心、不同模态的不同器官或病灶。尽管交互式分割方法只需要少量标注数据进行训练就可以用于多种场景的多种目标进行标注，但其也受到了在测试过程中需要人工交互的限制难以进一步减少医生的标注负担。此外，与本章内容相比，上章工作的主要不足还包括需要一部分密集标注的图像进行训练，而本章是为了解决稀疏标注情况下的训练问题。为了克服上述困难，本章主要从训练方法的角度出发，考虑如何利用弱监督学习和稀疏标注（例如，类别信息、点、线条或边界框）来训练全自动、高性能的分割模型来降低深度学习模型对数据依赖的同时减低数据的标注成本。

弱监督学习 (Weakly-supervised learning) 旨在利用稀疏的标注信息来训练模型减低数据标注成本的同时获得准确的结果。由于稀疏标注能为模型提供的监督信号有限，要实现上述目的存在两大挑战：1) 选择什么样的稀疏标注形式；2) 如何充分利用这些稀疏标注。针对上述挑战本研究调研发现，在深度学习时代之前，基于涂鸦标注 (scribble) 和传统机器学习的方法是最流行和通用的图像分割算法，例如 Graph Cuts^[2]、GrabCut^[65]、Random Walks^[72]、GrowCut^[73]、ITK-SNAP^[56]、Slic-Seg^[64]。图 2-13展示了密集标注和涂鸦标注之间的差异，可以发现相比密集标注涂鸦更容易获得且能很好的表示复杂的解剖学结构。此外，伪标签^[93]被广泛用于为未标记的图像/像素生成监督信号，其主要思想是利用部分标注信息来为无标注图像/像素产生高质量和可靠的伪标签，用于网络训练^[11,198]。但伪标签算法也受限于其固有的不足，即对同一幅图像模型倾向于输出与其伪标签一样的预测，而抵制对预测结果进行更新^[199]。

为了克服上述挑战，本章提出了一种基于双分支网络 (Dual-Branch Network, DBN) 和动态混合伪标签的训练策略 (Dynamically Mixed Pseudo Labels Supervision, DMPLS)，利用涂鸦标注训练高性能、低标注成本的医学图像分割模型。本章主要内容为：

- 1) 本章提出了一个双分支网络和动态混合伪标签策略，可利用稀疏的涂鸦标注训练分割模型。具体来说，该通过随机混合两个分支的输出生成高质量的伪标签，并使用生成的伪标签端到端地监督网络训练。

- 2) 在公开的心脏结构分割 MRI 数据集和腹部多器官分割 CT 数据集上进行了大量实验，证明了该方法的有效性。该方法在心脏结构分割和腹部多器官分割数据集上的性能均优于现有的涂鸦监督分割方法，并且也优于几种具有类似标注成本的半监督分割方法。

4.2 算法设计

本章提出的基于涂鸦监督的医学图像分割架构如图 4-1 所示。首先采用一个具有一个编码器和两个略有不同的解码器的网络来从涂鸦注释中学习，以分割目标对象。同时利用两个分支的输出生成硬的伪标签，用于辅助网络训练。值得注意的是，训练过程是端到端的，而不是多阶段^[88]或迭代修正策略^[87]。

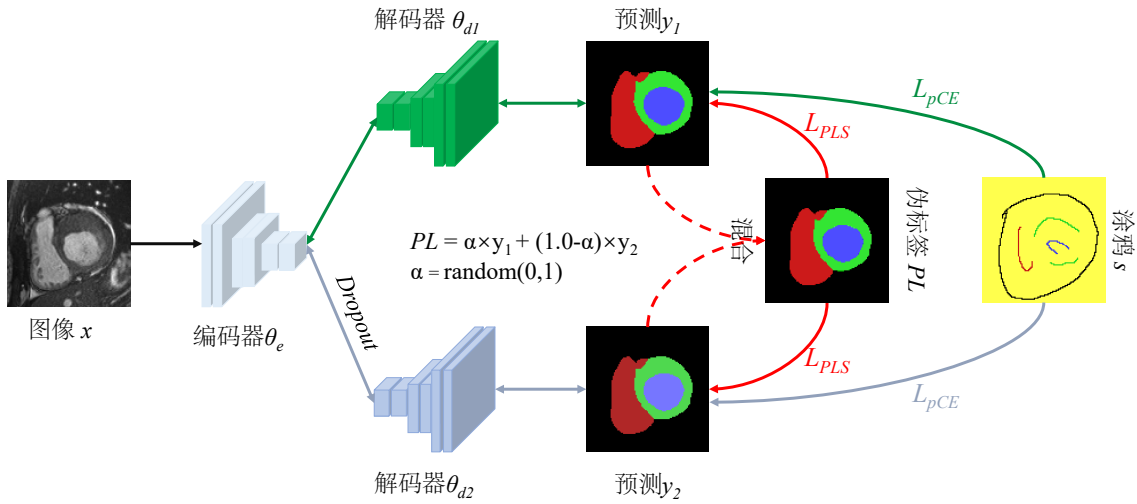


图 4-1 本章提出方法的概述。该方法由一个编码器 (θ_e)、主解码器 (θ_{d1}) 和辅助解码器 (θ_{d2}) 组成，并分别使用涂鸦注释进行训练 (L_{pCE})。同时，通过动态混合两个解码器的输出生成伪标签，并将其用作进一步网络训练的伪标签监督 (L_{PLS})

4.2.1 涂鸦监督学习

对于一般的涂鸦监督学习来说，数据集由配对的图像和涂鸦标注组成，其中涂鸦是具有类别或未知标签的一组像素。Can 等人^[88]使用交互式分割方法 Random Walks^[72]将涂鸦标注的像素传播到整个图像进行粗略分割结果，然后以全监督方式使用分割来训练深度网络。最近，出现了更好的替代方案^[7,90]，可直接使用涂鸦通过最小化部分交叉熵损失来训练分割模型：

$$L_{pCE}(y, s) = - \sum_c \sum_{i \in \omega_s} \log y_i^c \quad (4-1)$$

其中, s 表示二进制编码的涂鸦标注。 y_i^c 表示像素 i 属于类别 c 的预测概率。 ω_s 是涂鸦中的标记像素集合。与全监督的交叉熵损失函数相比, 该损失函数在模型训练过程中只考虑图像中有标注的像素及其类别忽略了未标注像素。因此该方法的监督信号较弱, 导致模型训练过程不稳定, 模型性能较差且模型容易过拟合到标注像素和类别。为解决此问题, 下面内容详细介绍了本章提出的方法细节。

4.2.2 双分支网络

本章所提出的网络结构 ($f(\theta_e, \theta_{d1}, \theta_{d2})$) 由一个共享的用于特征提取的编码器 (θ_e), 和两个独立且不同的用于分割和辅助训练的解码器组成 (θ_{d1}, θ_{d2}) (如图4-1)。具体来说, 本章将一个带扰动的解码器嵌入到了通用的 UNet^[8] 结构中作为辅助训练解码器, 在实现过程中使用了 Dropout^[158] 在特征级别引入扰动。这个设计有两个优点: (1) 它可以克服单分支网络中伪标签固有的弱点^[199], 因为两个分支的输出由于特征扰动而不同。(2) 它可以通过两个输出的集成生成伪标签, 但不需要训练两个网络, 并且编码器受益于两个单独的监督可以进一步提升特征提取能力^[10, 158]。值得指出的是, 一些最近的工作使用类似的架构进行一致性训练^[10, 158, 170] 或知识蒸馏^[9]。本研究在学习场景和监督策略上与已有方法存在许多显著差异。首先, Luo、Wu、Ouali 等人^[10, 158, 170] 专注于半监督学习, 而 Dolz 等人^[9] 则专注于知识蒸馏, 但章节的目标是基于涂鸦监督的医学图像分割。其次, 它们采用一致性正则化来监督网络, 而本章方法则通过随机混合两个输出来生成伪标签进行全监督学习。这些方法和目标的差异也产生了不同的训练方法、优化策略和实验结果。

4.2.3 动态混合伪标签

基于上一节提出的双分支网络, 本节进一步利用两个解码器的输出结果来促进模型训练的效果和稳定性。首先是通过动态混合两个预测结果来生成伪标签, 类似于 mixup^[200]。动态混合伪标签 (Pseudo Label, PL) 生成过程的数学定义如下:

$$PL = \operatorname{argmax}[\alpha \times y_1 + (1.0 - \alpha) \times y_2], \alpha = \operatorname{random}(0, 1) \quad (4-2)$$

其中, y_1 和 y_2 分别是主解码器和辅助解码器的输出。 $\alpha \sim U(0, 1)$ 是随机混合系数, 每次迭代都从区间 $[0, 1]$ 的均匀分布中随机采样生成。例如, 当 $\alpha = 0.0$ 和 $\alpha = 1.0$ 时, PL 分别变为 p_2 和 p_1 。当 $\alpha = 0.5$ 时, PL 是 p_1 和 p_2 的平均值。该策略可以提高伪标签的多样性, 并避免了伪标签策略的固有弱点 (模型倾向于输出与伪标签一样的预测结果, 而不随着训练进行持续更新^[199])。 argmax 为最大值的参数用于生成伪标签。与一致性正则化相比^[10, 158], 这种策略切断了 θ_{d1} 和 θ_{d2} 之间的梯

度，以保持它们的独立性，而不是直接强制一致性。基于这种设计，监督信号通过动态混合伪标签策略从少数涂鸦标注像素扩展到整个图像。然后，本章进一步使用生成的 PL 来分别监督 θ_{d1} 和 θ_{d2} 以辅助网络训练。提出的伪标签监督（Pseudo Label Supervision, PLS ）被定义为：

$$L_{PLS}(PL, y_1, y_2) = 0.5 \times (L_{Dice}(PL, y_1) + L_{Dice}(PL, y_2)) \quad (4-3)$$

其中 L_{Dice} 是广泛使用的 Dice 损失函数，也可以用交叉熵损失或其他分割损失函数替代。

4.2.4 目标函数

本节提出方法的训练流程伪代码如算法4-1所示，对于图像中涂鸦标注了的像素，本章用部分交叉熵损失函数（公式4-1）对两个预测分支进行分别独立监督，该损失在梯度反向传播期间只考虑标记的像素而忽略了未标记的像素。对于没有标注的像素，本章使用了上一节提出的伪标签监督损失来进行模型训练（公式4-3）。最后，本章所提出的双分支网络通过最小化此联合目标函数来利用涂鸦标注进行训练：

$$L_{total} = \underbrace{0.5 \times (L_{pCE}(y_1, s) + L_{pCE}(y_2, s))}_{\text{涂鸦监督}} + \lambda \times \underbrace{L_{PLS}(PL, y_1, y_2)}_{\text{伪标签监督}} \quad (4-4)$$

其中 λ 是一个权重系数，用于平衡部分交叉损失和伪标签监督损失。

4.3 实验验证

4.3.1 实验数据与实现细节

为了公平的验证和比较所提出的方法，本章在两个广泛使用的公开涂鸦标注数据集上评估了本章提出的方法，并将其与几种最先进的方法进行了比较。

1) 心脏结构分割 MRI 数据集： ACDC 数据集^[174] 包含了不同磁场强度和空间分辨率的 100 位患者的 200 个短轴心脏 MRI 图像。每个图像包含 6 到 18 张不等的 2D 切片，这些切片的长宽范围分别为 154-276 和 174-384。临床医生在每位患者的舒张末期和收缩末期 MRI 图像上标注了右心室（RV）、心肌（Myo）和左心室（LV）结构。此外，Valvano 等人^[92] 在他们的既往研究中提供了每张图像的涂鸦标注，如图2-13所示。基于此数据，本章采用了五折交叉验证的方法评估了提出的方法和其他最先进的方法，其中每折中 80 位患者数据用于训练，剩下的 20 位患者数据用于测试。考虑该数据集的层间距较大（大都超过了 10 mm），本章使用

算法 4-1 基于双分支网络和动态混合伪标签监督方法的训练过程

Data: 来自不同数据集的图像及涂鸦标注对 (x^k, s^k) , 双分支网络 $f(\theta_e, \theta_{d1}, \theta_{d2})$

Result: 训练得到的图像分割模型

```

1 开始训练;
2 while 模型未收敛或训练未停止 do
3   获得输入图像的不同分支输出  $y1^k$  和  $y2^k$ ;
4   根据不同分支输出  $y1^k$  和  $y2^k$  和公式4-2, 计算出动态混合伪标签  $PL^k$ ;
5   根据公式4-1分别计算两个分支的涂鸦监督损失;
6   根据公式4-3分别计算两个分支的伪标签监督损失;
7   根据公式4-4计算模型梯度, 更新模型参数;
8 end

```

了 2D 网络进行逐层分割, 并将结果堆叠成 3D 进行指标评估。

2) 腹部器官分割 CT 数据集: **WORD**^[44] 数据集包含了来自单个中心的 150 名腹部疾病患者的放射前 CT 图像。每个 CT 图像由 159-330 张 512×512 的切片组成, 这些 CT 图像的层内分辨率为 $0.976 \text{ mm} \times 0.976 \text{ mm}$, 层间分辨率为 $2.5\text{-}3.0 \text{ mm}$ 。本章的目标是从这些腹部 CT 中分割以下七个器官: 肝脏、脾脏、左肾、右肾、胃、胆囊和胰腺。图2-12展示了 **WORD** 数据集的密集和涂鸦标注, 器官和背景的涂鸦是由 Luo 等人^[44] 在轴向视图中通过脚本模仿医生勾画生成的。该数据集被官方化分为 100、20 和 30 个 CT 图像分别用于模型的训练、验证和测试。

3) 实验参数设置: 对 ACDC 数据集来说, 预处理阶段本章将每个切片的灰度值被归一到 $[0,1]$ 。实验过程中 2D UNet^[8] 被用作主干网络, 并为其添加了一个辅助解码器以扩展为双分支网络, 此外还在辅助解码器之前的瓶颈处加入了一个 dropout 层 (比例为 0.5) 以引入扰动。在模型训练过程中网络训练, 随机旋转和翻转等操作用来做数据增强, 此外还将增强后的图像调整为 256×256 的固定尺寸作为模型输入。训练中的批量大小和总迭代次数被分别设置为 24 和 30,000 (≈ 527 轮)。测试过程中, 双分支网络为逐个切片生成预测, 然后将它们堆叠成 3D 图像。在 **WORD** 数据集上, 本章采用了一个固定的窗宽/窗位 (400/50) 以更好的展示腹部器官, 并随后将 CT 图像的 Hounsfield 数值归一化到 $[0,1]$ 范围内。此外, 本研究使用 3D-UNet^[32] 作为 3D 主干网络, 并将其扩展为 3D 双分支网络。在训练期间, 本研究仅通过随机裁剪进行数据增强, 随机裁剪的补丁大小为 $80 \times 96 \times 96$ 。由于内存限制, 实验的批量大小被设置为 1, 最大迭代次数被设为 60k (≈ 600 轮)。最终分割结果通过滑动窗口策略获得。

本实验中方法均由 PyTorch1.11^[189] 实现, 并在一个带有 8 块 NVIDIA GTX1080TI

GPUs 的 Ubuntu20.04 的服务器上进行训练和测试。在训练阶段，所有方法都采用了随机梯度下降（SGD）优化器进行模型训练，其中 SGD 的初始学习率 (learning rate, lr) 为 0.01，动量为 0.9，权重衰减系数为 10^{-4} 。此外，本实验还采用了多项式学习率调度策略来用到调整学习率，具体策略如下： t 轮迭代的学习率为 $lr_t = (1.0 - t/t_{max})^{0.9} \times lr$ ，其中 t_{max} 指的是最大迭代次数。为了进行公平的算法比较，本实验在两个数据集上分别采用了 Valvano 等人^[92] 和 Luo 等人^[44] 的实验参数设定；也继续沿用了他们的评价指标 Dice (%) 和 95% 豪斯多夫距离（95% Hausdorff Distance, 95HD）。

4.3.2 算法性能比较

1) 心脏结构分割 MRI 数据集：为了证明所提出方法的有效性，本章首先在基于公开的心脏结构分割 MRI 数据集进行了对比实验。在比较不同训练策略对模型分割结果影响的实验中，本章将“pCE”^[87] 作为所有方法的比较基准即所有方法的结果下限（lower bound），该方法在模型迭代过程中只计算涂鸦标注像素的梯度并进行模型更新并忽略全部未标注像素。同时也展示了基于密集标注的全监督模型实验结果即所有方法的结果上限“FullSup”（upper bound）。除此之外，实验还将提出的 DBN 网络和 DMPLS 策略与现有的几种最先进的弱监督分割方法进行了比较，这些方法包括 1) 基于 Random Walks 生成伪标签的 RW^[72]；2) 基于不确定性教师学生网络和变换一致性的 USTM^[91]；3) 基于伪标签更新的 Scribble2Label^[90]；4) 基于活动轮廓的 MLoss^[89]；5) 基于熵最小化的 EM^[201]；6) 基于条件随机场的正则化损失 RLoss^[7]。表 4-1 的第一部分列出了本章提出方法与七种现有的弱监督学习方法在五折交叉验证情况下的定量结果比较。可以发现，提出方法在平均 Dice 方面表现最佳（ p 值 < 0.05 ），并且在 95HD 指标上位列第二，优于大多数其他方法。随后，本章进一步将所提出方法与其他流行的低成本医学图像分割方法进行了比较，例如半监督学习方法。Dorent 等人^[202] 的研究结果表明，在不同标注成本要求下，不同标注方式和模型训练策略得到的实验结果显著不同，因为本章进一步研究了在使用非常相似的标注成本时这些方法的性能差异。为此，本章探究了分别以弱监督和半监督学习两种策略来训练网络的分割结果。对半监督学习来说，本实验使用了 10% 的训练集（8 个患者的 16 个 MRI 数据）作为有标签数据，其余作为无标签数据，因为这种实验设置和获得全部图像的涂鸦标注需要差不多的标注成本^[92,203]。对于部分监督（PS）学习来说（半监督方法的下限），本实验只使用 10% 的标记数据来训练网络。对于半监督学习，10% 的标记数据和 90% 的无标记数据结合起来共同用于模型训练。为了公平对比，本章采用了四种广泛

使用的半监督分割方法进行性能比较：1) 深度对抗网络 (DAN) [100]，2) 对抗熵最小化 (AdvEnt) [204]，3) 平均教师模型 (MT) [94]，4) 基于不确定性感知的平均教师模型 (UAMT) [97]。Dice 和 95HD 的定量比较结果如表 4-1 中第二部分所示。这些结果表明，涂鸦注释在具有类似标注预算时可以获得比像素级注释更好的结果 (Ours 与 PS 比较)。此外，提出的弱监督方法在心脏磁共振分割中明显优于现有的半监督方法。最后，本章还在表 4-1 的最后一行展示了心脏结构分割模型的性能上限 (FullSup)。实验结果可以发现，与像素级标注相比，本章方法略逊色于完全监督学习。但其需要比像素级注释更少的注释成本 (可节约 90% 左右的标注成本)。图 4-2 展示了 ACDC 数据集上使用现有方法和提出方法获得的分割结果，以及相应的金标准 (病历号为 patient026_frame01)。可以观察到，提出方法得到的结果与金标准更相似。这进一步表明，涂鸦是一种潜在的可以降低注释成本的数据标注方法。下一节将在另一个公开的腹部器官分割 CT 数据集上进一步验证所提出方法在处理涂鸦标注问题时的有效性。

2) 腹部器官 CT 数据集：与上一节实验设置类似，本节在另一个公开数据集上验证了所提出方法的有效性。不同方法在 WORD 数据集上分割结果的 Dice 和 95HD 指标如表 4-2 所示。从表 4-2 中可以观察到，实验下限 pCE 分割结果的平均 Dice 为 75.46%。USTM [91] 和 S2L [90] 分割结果的平均 Dice 分别为 78.85% 和 78.39%。RLoss [7] 在现有弱监督方法中获得了最高的 Dice (85.12%)。此外，可以观察到本章方法的实验结果优于 RLoss [7]，平均 Dice 和 95HD 分别为 86.79% 和 8.85mm。此外，还可以发现本章所提出的方法在大多数器官上在 Dice 和 95HD 两个评价指标上均显著优于最佳现有方法 RLoss [7]，特别是在结构复杂和体积小的器官上，如胃、胆囊和胰腺等。图 4-3 展示了这些方法之间的可视化比较 (结果来自病人 word_0077)。从这张图上可以看出，pCE 获得的分割结果存在大量的假阳预测，EM [201] 和 RLoss [7] 在三维空间中的分割结果也存在一些假阳和欠分割。与这些方法相比，本章方法取得了更准确的分割结果，并进一步减少了过/欠分割区域。

4.3.3 消融实验

为了更好的验证提出方法的主要组成部分的性能贡献，本小节在心脏结构 MRI 数据集上进行了详情的消融实验和结果分析。在消融实验中，首先评估了目标函数 4-4 中 λ 的敏感性。因为 *PLS* 项在所提出的框架中起着至关重要的作用，控制了网络训练过程中伪标签的使用。在本实验中 λ 分别被设置为 {0.01, 0.1, 0.2, 0.3, 0.5, 1.0} 进行模型训练，并评估不同 λ 训练的模型获得结果的性能。图 4-4 展示了不同 λ 设置下基于 5 折交叉验证的 RV、Myo、LV 的分割结果变

表 4-1 本章提出方法在 ACDC 数据集上与现有弱监督/半监督方法在 DSC 和 95HD (mm) 的结果比较。所有结果均基于相同骨干结构 (2D UNet) 的 5 折交叉验证。* 表示本章方法与第二名方法 (RLoss [7]) 配对 t 检验时 p 值 < 0.05。WSL, SSL 和 FSL 分别表示弱监督学习, 半监督学习和全监督学习

	RV		Myo		LV		平均		
	Dice (%)	95HD (mm)	Dice (%)	95HD (mm)	Dice (%)	95HD (mm)	Dice (%)	95HD (mm)	
WSL	pCE [87]	62.50±16	187.2±35.2	66.82±9.51	165.1±34.4	76.61±15.62	167.7±55.0	68.68±13.72	173.3±41.5
	RW [72]	81.32±11.33	11.1±17.3	70.81±6.62	9.8±8.9	84.42±9.12	9.2±13.0	78.78±9.12	10.0±13.1
	USTM [91]	81.50±11.51	54.7±65.7	75.63±8.1	112.2±54.1	78.53±16.18	139.6±57.7	78.62±11.93	102.2±59.2
	S2L [90]	83.31±10.3	14.6±30.9	80.58±6.91	37.1±49.4	85.56±12.11	65.2±65.1	83.22±9.81	38.9±48.5
	MLoss [89]	80.88±9.27	17.1±30.8	83.18±5.45	28.2±43.2	87.61±9.28	37.9±59.6	83.89±7.8	27.7±44.5
	EM [201]	83.89±10.78	25.7±44.5	81.18±6.22	47.4±50.6	88.67±9.9	43.8±57.6	84.56±8.89	39.0±50.9
	RLoss [7]	85.56±10.09	7.9±12.6	81.67±5.43	6.0±6.9	89.56±8.56	7.0±13.5	85.56±7.96	6.9±11.0
	Ours	86.11±9.62	7.9±12.5	84.21±5.42*	9.7±23.2	91.28±8.21*	12.1±27.2	87.22±7.7*	9.9±21.0
	PS [8]	65.89±26.12	26.8±30.4	72.38±17.64	16.0±21.6	79.02±20.51	24.5±30.4	72.43±21.42	22.5±27.5
	DAN [100]	63.89±26.02	20.6±21.4	76.42±14.36	9.4±12.4	82.54±18.56	15.9±20.8	74.32±19.67	15.3±18.2
SSL	AdvEnt [204]	61.46±29.62	20.2±19.4	76.04±15.11	8.5±8.3	84.78±15.89	11.7±18.1	74.11±20.22	13.5±15.3
	MT [94]	65.33±27.11	18.6±22.0	78.45±0.118	11.4±17.0	84.56±15.33	19.0±26.7	76.11±18.01	16.3±21.9
	UAMT [97]	66.03±26.72	22.3±22.9	77.32±12.91	10.3±14.8	84.67±15.67	17.1±23.9	76.02±18.45	16.6±20.5
	FullSup [8]	88.22±9.51	6.9±10.8	88.32±4.22	5.9±15.2	92.99±7.42	8.1±20.9	89.82±7.02	7.0±15.6

表 4-2 不同弱监督分割方法在 WORD 数据集上的 Dice(%) 和 95HD (mm) 指标的定量比较。所有结果均基于相同骨干结构 (3D-UNet)。* 表示本章方法与第二名方法 (RLoss^[7]) 配对 t 检验时 p 值 <0.05

	肝脏	脾脏	左肾	右肾	胃	胆囊	胰腺	平均
Dice (%)								
pCE (lower bound)	86.53±3.61	86.81±5.75	86.25±4.46	89.41±2.97	61.09±12.20	56.61±20.12	61.55±10.45	75.46±8.51
USTM ^[91]	80.57±3.87	86.49±4.71	87.15±4.21	89.26±2.35	77.33±6.19	63.94±17.17	67.24±9.49	78.85±6.86
S2L ^[90]	68.42±3.51	89.98±5.51	88.10±4.05	89.90±4.05	81.39±6.07	60.99±19.55	69.98±8.49	78.39±7.32
EM ^[201]	89.60±1.72	87.76±5.65	87.29±3.88	81.32±4.79	77.74±7.34	65.83±16.80	73.52±7.54	80.44±6.82
RLoss ^[7]	93.54±1.05	93.29±2.52	90.67±3.16	92.51±3.25	83.51±4.58	69.61±15.02	72.68±8.57	85.12±5.45
Ours	94.40±0.91 *	92.24±2.68	92.37±2.37 *	92.96±1.73	87.91±3.79 *	71.14±15.37	76.49±7.85 *	86.79±4.84 *
FullySup (upper bound)	96.37±0.74	95.42±1.55	94.95±1.58	95.33±1.34	90.08±4.42	75.33±13.21	80.90±7.67	89.77±4.36
95HD (mm)								
pCE (lower bound)	64.97±33.39	59.65±46.21	63.51±46.81	56.12±50.34	130.05±30.53	69.23±36.97	99.41±13.35	77.56±36.80
USTM ^[91]	99.07±28.83	13.89±18.13	30.58±6.15	40.81±49.34	90.49±44.52	61.75±52.17	93.42±14.41	61.43±34.79
S2L ^[90]	48.40±14.44	6.70±6.04	6.49±1.94	6.02±2.43	38.81±41.89	26.52±29.29	34.62±29.01	23.94±17.86
EM ^[201]	32.27±24.35	24.79±39.25	7.32±3.51	121.68±10.83	107.03±54.52	29.62±45.69	33.57±31.29	50.90±29.92
RLoss ^[7]	7.29±3.13	7.24±18.11	7.31±2.21	4.39±2.00	14.58±8.55	11.79±16.43	16.66±23.68	9.90±10.59
Ours	5.57±2.62 *	9.42±2.79	4.42±2.03 *	3.58±1.11 *	9.99±6.26 *	17.13±22.89	11.99±16.28	8.85±11.30
FullySup (upper bound)	3.45±1.40	3.01±3.31	2.84±1.05	2.47±0.71	10.01±8.36	10.50±15.80	7.71±6.86	5.71±5.36

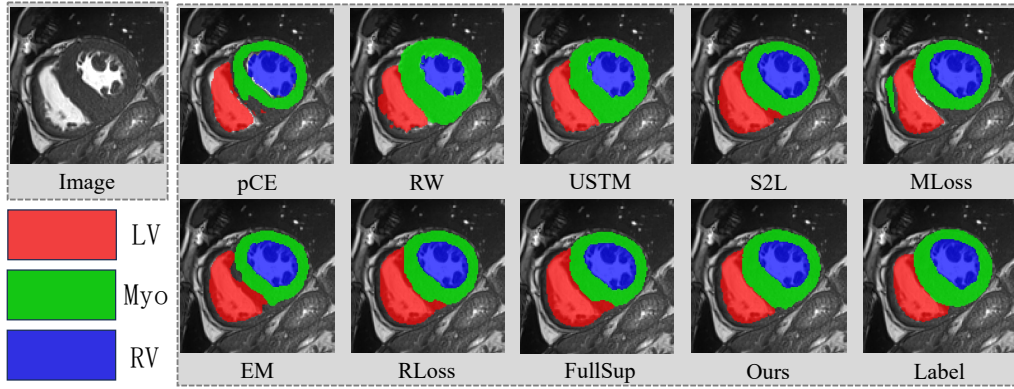


图 4-2 提出方法与其中现有七种弱监督学习方法在心脏结构分割上的可视化比较

化曲线以及它们的平均结果。从图上可以观察到， λ 从 0.01 增加到 0.5，模型性能在 Dice 和 95HD 方面都获得了提高。当 λ 被设置为 1.0 时，与 0.5 相比，分割结果仅略微下降（平均 Dice 为分别为 87.2% 和 87.0%）。这些实验结果表明，本章所提出的方法对 λ 不敏感。

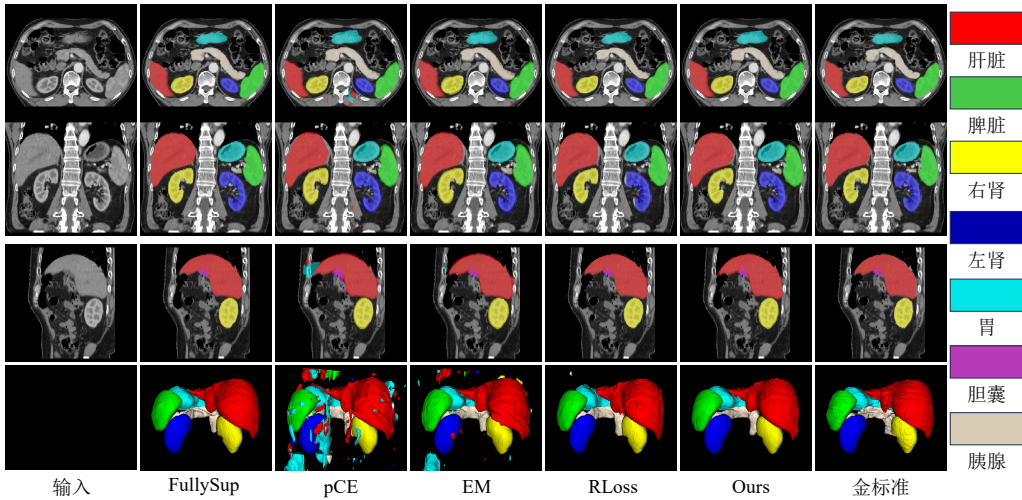
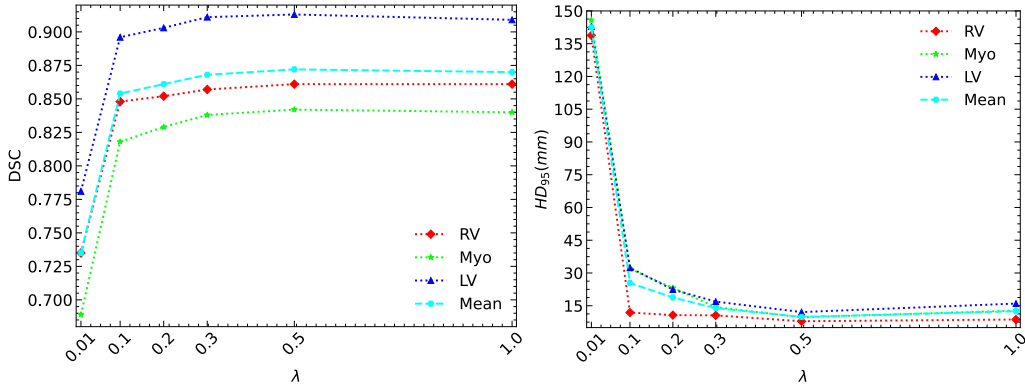


图 4-3 本章方法与其他弱监督方法在 WORD 数据集上分割结果的可视化比较

消融实验进一步研究了双分支网络使用不同监督方法的分割结果：1）一致性正则化 (Consistency Regularization, CR)^[9]，它鼓励两个分支的预测结果变的相似；2）交叉伪监督 (Cross Pseudo Supervision, CPS)^[10,11]，它使用一个分支的输出作为伪标签来监督另一个分支；3）所提出的动态混合伪标签方法，动态混合两个输出以为两个解码器分别训练生成伪标签。本实验使用涂鸦标注和上述三种监督策略对双分支网络进行训练。量化评估结果如表 4-3 所示。可以观察到，与 CR 和 CPS

图 4-4 模型在不同 λ 下的分割结果曲线

相比, 使用本章提出的 *PLS* 训练的模型可以取得最佳性能。此外, 本实验还调查了 α 被设置为固定值 (0.5) 和动态值时的模型性能。实验结果表明了所提出的动态混合策略的有效性。此外, 本实验还发现主要解码器分支 (θ_{d1}) 和辅助解码器分支 (θ_{d2}) 取得了非常相似的结果。

4.4 讨论与分析

卷积神经网络 (CNN) 和 Transformer 在自动医学图像分割方面取得了令人鼓舞的进展 [8, 34, 141]。大多数基于深度学习的方法依赖大规模的精细标注图像进行训练模型。然而, 收集大规模且精细标注的医学图像数据集仍然是一项昂贵且耗时的工作, 因为它需要领域知识和临床经验 [170, 205]。最近, 许多工作致力于减少模型训练的标注成本以缓解这个问题。例如, 半监督学习 (SSL) 将少量标记数据与大量未标记数据结合起来进行网络训练 [95, 153, 170, 206]。弱监督学习 (WSL) 使用稀疏标注来进行模型训练, 而不是逐像素标注 [9, 92, 202]。考虑到收集稀疏标注 (点、涂鸦和边界框) 相对于像素级标注更容易 [87], 而涂鸦在标注复杂器官或病灶时比边界框和点具有更好的通用性 [87, 92] (如图 2-13 和 2-12 所示), 本章侧重于探索利用涂鸦标注来有效和稳健地训练高性能的医学图像分割网络。

该项工作主要是从网络结构和监督信号设计两个方面入手, 开发基于涂鸦标注的全自动高性能医学图像分割模型。首先是双分支网络, 本工作利用特征扰动来提高两个分支的预测差异, 并为两个分支设计独立的监督信号进一步保证两个分支预测的独立性。消融实验结果也可以发现双分支设计能够取得比单分支网络更好结果, 其在测试阶段并没有增加任何计算开销。其次是基于双分支预测的动态混合伪标签, 此设计缓解了传统伪标签算法存在的固有内在缺陷, 即模型从固定伪标签中学习容易过拟合, 而导致更新后的模型输出和原伪标签相似。消融实验结果也证明了, 此方法通过动态融合的策略来生成更多样性的伪标签有利于模

表 4-3 双分支网络对不同监督策略的消融实验研究。Single 表示仅使用 pCE 的 UNet^[8]。CR 表示不同解码器分支之间的一致性正则化^[9]。CPS 是 Chen 和 Wu 等人^[10, 11]提出的交叉伪监督策略。Ours 是本章提出的 PLS 策略。 θ_{d1} 和 θ_{d2} 分别表示主要和辅助解码器分支的预测

	RV		Myo		LV		平均	
	Dice (%)	95HD (mm)	Dice (%)	95HD (mm)	Dice (%)	95HD (mm)	Dice (%)	95HD (mm)
Single ^[87]	62.5±16	187.2±35.2	66.8±9.5	165.1±34.4	76.6±15.6	167.7±55.0	68.6±13.7	173.3±41.5
Dual+CR ^[9]	84.4±10.6	20.1±37.2	79.8±7.0	62.2±55.7	87.3±10.1	63.4±65.5	83.8±9.2	48.6±52.8
Dual+CPS ^[10, 11]	84.9±9.9	12.4±25.6	83.3±5.6	19.3±33.5	90.5±9.1	18.3±35.8	86.3±8.2	16.6±31.6
Ours ($\alpha=0.5, \theta_{d1}$)	85.5±10.1	8.6±13.9	83.7±5.3	13.6±29.1	90.8±8.6	15.8±34.1	86.6±8.0	12.6±25.7
Ours ($\alpha=\text{random}, \theta_{d1}$)	86.1±9.6	7.9±12.5	84.2±5.4	9.7±23.2	91.3±8.2	12.1±27.2	87.2±7.7	9.9±21.0
Ours ($\alpha=\text{random}, \theta_{d2}$)	86.1±9.8	7.3±10.3	84.0±5.8	10.9±24.5	91.1±8.6	11.3±26.4	87.1±8.0	9.8±20.4

型的进一步学习并生成更准确的预测。此外该策略也显著好于经典的一致性正则化（CR）和交叉伪标签（CPS），为伪标签生成提供了一种新的思路。

从分割方法的通用性来说，双分支网络和动态混合伪标签策略是一种即插即用的学习策略而不依赖任何特殊方法设计。从模型结构设计的方面来说，任何常见的分割模型均可在很小的改动下扩展成双分支用于模型设计。从分割任务上来说，该方法不需要针对特定任务进行调整或重新设计即可取得较好的实验结果，心脏结构 MRI 和腹部器官 CT 数据集上的优越性能均证明了该方法的在 2D 和 3D 分割任务上的通用性。从方法的鲁棒性来说，本方法只有一个超参数 λ 且消融实验表明模型性能对 λ 的数值设置不敏感，证明了该方法的鲁棒性。所以，可以认为该方法是一种可利用涂鸦标注训练高性能分割模型的有效方法。最近，Wu 等人^[10]证明了该方法在医学图像半监督分割任务上也可以取得优越的实验结果，也侧面说明了该方法有潜力成为低成本标注分割的统一训练策略。

4.5 本章总结

本章针对低标注成本的医学图像分割问题，提出了一种基于双分支网络和动态融合伪标签监督策略利用涂鸦标注训练高性能低成本的医学图像分割模型。具体来说，本章首先提出了一种双分支网络来以端到端的方式从涂鸦标注进行学习。基于双分支网络，本章进一步提出了一种动态混合伪标签策略，将涂鸦标注从极少的标注像素中传播到整个图像并进一步监督网络训练。最后，在两个公开的心脏结构分割 MRI（ACDC）和腹部器官 CT 分割数据集（WORD）上的实验验证了该方法的有效性。首先在使用相同的涂鸦标注情况下，该方法胜过了七种最近的涂鸦辅助分割方法；其次在相同标注成本的情况下，该方法也优于四种最近的半监督方法方法。在未来，该方法可以扩展以处理不同模态、器官和稀疏标注类型，并评估其在这些有挑战性的医学图像分割任务上的性能。

第五章 基于双任务一致性的半监督医学图像分割

5.1 引言

上一章设计了双分支网络和动态混合伪标签策略以利用涂鸦标注训练医学图像分割模型，从而降低医学图像标注成本。通过两个公开数据集的实验表明双分支网络可以生成更具有多样性的预测，进一步动态混合伪标签策略利用双分支网络的多样性预测生成新的混合伪标签为整幅图像提供监督信号进而提高模型的分割性能。该工作为基于有限标注数据的医学图像分割任务提供了有效的解决方法，即利用动态伪标签策略和涂鸦标注为未标注像素生成高质量伪标签，并利用这些高质量伪标签和稀疏标注直接训练分割模型。尽管涂鸦标注的弱监督学习可以获得较好的实验结果，但涂鸦标注难以准确描述目标对象的解剖学结构。此外，仅标注整个数据集的一小部分也可以大大降低标注成本，如何利用小部分标注数据和大量未标注数据训练模型也受到了广泛关注。因此，本章旨在探索如何利用少量标注数据和大量未标注数据结合半监督学习（SSL）提高模型的分割性能。

在本章中，提出了一种基于双任务一致性的半监督医学图像分割算法（Dual-Task Consistency, DTC）。该方法的理念在于鼓励模型从全局层面水平集回归任务与像素层面分类任务的一致性约束中学习，同时引入几何约束并充分利用大量未标记的数据进行模型训练。具体而言，该方法主要包含了三个主要组成部分。首先是像素和水平集预测模型（双任务分割网络），本章将分割问题定义为两个不同的任务，即局部像素级预测和回归水平集函数回归，其中零水平集即为分割轮廓。本方法受到 Ma 等人和 Xue 等人工作^[207-209]的启发，使用双分支模型来同时预测水平集函数和像素分类概率来提高模型的性能。其次是可微的跨任务转换层，本方法使用平滑的海维赛德函数^[209]将水平集函数转换为分割预测概率图。最后是一个联合的损失函数进行半监督学习，即同时进行监督学习和无监督学习，本章提出了一个跨任务一致性约束来鼓励模型的像素级分割预测和水平集函数转换而来的预测保持一致。这种设计不仅可以增强全监督学习的性能，还可以有效利用未标记的数据进行无监督学习。值得注意的是，SASSNet^[171]也提出了一个包含像素级预测分支和形状感知回归分支的多任务网络用于半监督医学图像分割。对于半监督学习来说，SASSNet 只使用形状分支并结合对抗正则化建立一致约束来平滑整个未标注数据集的预测分布，但是该策略仍然属于数据层面的正则化，没有考虑不同任务之间的差异。此外，本工作观察到在模型训练过程中来自不同任务分支的特征信息可以相互补充，而且不同的任务差异也可能会带来潜在的模型

扰动。然后，这项工作中提出了最重要的问题：是否可以显式地建立完全不同于数据一致性正则化的多任务一致性正则化？显然，答案是肯定的，建立条件是不同的任务分支的输出可以被映射或转换到相同预测空间，这样就可以直接计算两个任务之间的一致性用于训练。总的来说，本章针对半监督医学图像分割方法提出了一种简单且高效的双任务一致性约束，通过迫使不同任务预测保持一致来利用未标记的数据。本工作在半监督学习中首次引入了任务一致性正则化的概念，并提出了一个多任务一致性算法，本工作的主要贡献为：

- 1) 在全监督学习中，本章所提出的双任务一致性正则化表现优于不同任务之间的单独和联合监督。
- 2) 在半监督学习中，本章所提出的方法在几个临床数据集上优于最先进的半监督医学图像分割算法。
- 3) 与既往工作相比，本章所提出的方法在训练开销和时间方面的成本更少。此外，该方法是一个通用的半监督医学图像分割方法，只要不同任务之间的存在可微分变换就可以直接适用。

5.2 算法设计

本节中介绍了基于双任务一致性的半监督医学图像分割算法。总体框架如图 5-1 所示，该双分支网络对于每个输入的三维医学图像产生一个水平集函数预测和像素级分类预测。在模型训练过程中，对有标注图像，像素级分类分支由标注直接约束，水平集函数预测分支由标注产生的水平集函数约束。对于无标注数据，本章提出了一种新的双任务一致性约束，通过最小化预测的逐像素概率图和水平集函数之间的差异来从未标记的数据中学习。为了建立一致性，本章使用了一个可微的转换层将水平集函数转换为预测概率图，该转换方程由平滑的海维赛德函数实现。在接下来的两个小节中，本文首先介绍了双任务一致性策略，然后介绍了通过双任务一致性进行分割的半监督训练。

5.2.1 双任务一致性

在通用半监督学习中，一致性损失被精心设计用于在数据层面上促进预测变的平滑，即同一数据在不同变换^[155]和扰动^[158]下模型预测应该相同。与数据一致性不同，本工作强制在像素级分类任务（定义为任务 1）和全局水平集回归任务（定义为任务 2）之间实现任务级别的一致性。在现有的研究中，将分割任务定义为像素级分类任务已经被广泛研究，而水平集表示^[60]是一个传统的分割表示，这种表示方法可以更好的建模分割目标的几何轮廓和距离信息，最近在与 CNN^[140]

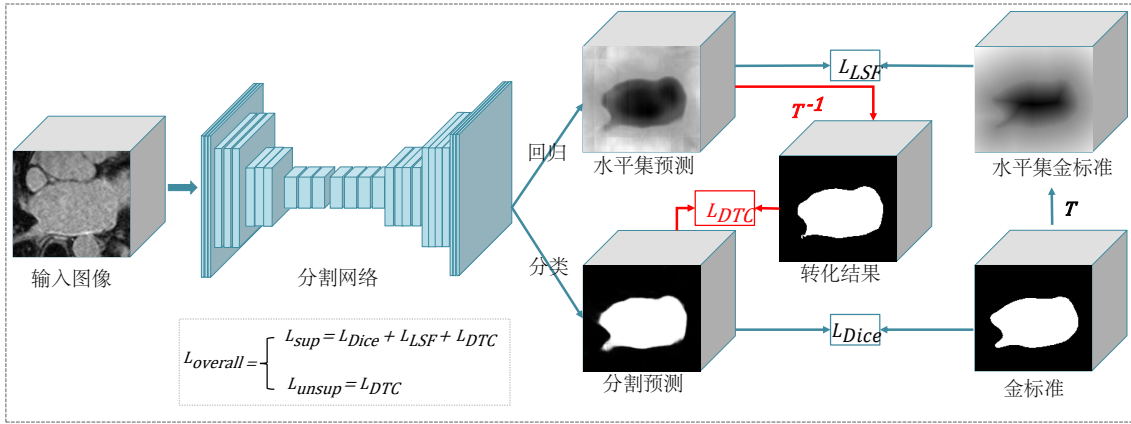


图 5-1 用于半监督医学图像分割的双任务一致性算法流程图。该网络由像素级分类分支（任务 1）和水平集函数回归分支（任务 2）组成，采用了广泛使用的编码器-解码器网络作为分割网络，即 VNet^[5,6]。该模型通过最小化有标注数据上的全监督损失函数 L_{Dice} 、 L_{LSF} 和无标注数据以及有标注数据上的双任务一致性损失 L_{DTC} 进行模型优化。 T 函数用于金标准转换成水平集表示以进行监督训练。 T^{-1} 函数将水平集表示转换为分割结果，以计算双任务一致性损失 L_{DTC}

相结合后重新获得了大量的关注和研究。本章将水平集表示定义如下：

$$\mathcal{T}(x) = \begin{cases} -\inf_{y \in \partial S} \|x - y\|_2, & x \in \mathcal{S}_{in} \\ 0, & x \in \partial S \\ +\inf_{y \in \partial S} \|x - y\|_2, & x \in \mathcal{S}_{out} \end{cases} \quad (5-1)$$

其中， x ， y 分别表示分割结果中的两个不同像素/体素， ∂S 是零水平集，也代表分割目标的轮廓。 $\mathcal{S}_{in}$ 和 $\mathcal{S}_{out}$ 分别表示分割目标的内部区域和外部区域。然后，将 $\mathcal{T}(x)$ 定义为从金标准到水平集函数地图的任务转换，如公式 5-1 所示。在将 LSF 任务的输出映射到分割输出空间时，自然而然会考虑使用 $\mathcal{T}(x)$ 的逆变换。然而，由于其逆变换不可微，在训练中使用 $\mathcal{T}(x)$ 的准确逆变换是不切实际的。因此，本章采用了一个平滑逼近来近似水平集函数的逆变换，假设希望在转换后的预测结果中， $\mathcal{S}_{in}$ 的值被分配为 1，而 $\mathcal{S}_{out}$ 的值被分配为 0，该平滑逼近定义为：

$$\mathcal{T}^{-1}(z) = \frac{1}{1 + e^{-k \cdot z}} = \sigma(k \cdot z) \quad (5-2)$$

其中, z 表示像素/体素 x 处的水平集数值。 $\mathcal{T}^{-1}(z)$ 的表述既精细又简单, 因为它可以被等效为 k 倍输入的 sigmoid 函数, 其中 k 需要被设置的尽可能大, 以逼近 $\mathcal{T}(x)$ 的逆变换。因此, $\mathcal{T}^{-1}(z)$ 可以很容易地实现为 k 倍任务 2 的输出 sigmoid 激活函数。此外, 其可微性证明如下所示:

$$\begin{aligned}\frac{\partial \mathcal{T}^{-1}}{\partial z} &= \left(\frac{1}{1 + e^{-kz}} \right)' \\ &= k \cdot \frac{1}{1 + e^{-kz}} \cdot \left(1 - \frac{1}{1 + e^{-kz}} \right)\end{aligned}\quad (5-3)$$

虽然这样一个近似的转换函数将使得任务 2 的预测空间对齐到与任务 1 的预测空间相同, 但它依然保留了不同任务层面的预测差异, 因为任务 1 关注像素级分类, 而任务 2 则更关注几何位置等结构信息回归。因此, 对于来自数据集 $\mathcal{D}$ 的输入 $\mathbf{X}$, 定义了双任务一致性损失 $\mathcal{L}_{DTC}$, 以强制任务 1 的预测 $f_1(\mathbf{x}_i)$ 与任务 2 预测的转换结果 $\mathcal{T}^{-1}(f_2(\mathbf{x}_i))$ 之间的一致性:

$$\begin{aligned}\mathcal{L}_{DTC}(\mathbf{x}) &= \sum_{\mathbf{x}_i \in \mathcal{D}} \left\| f_1(\mathbf{x}_i) - \mathcal{T}^{-1}(f_2(\mathbf{x}_i)) \right\|^2 \\ &= \sum_{\mathbf{x}_i \in \mathcal{D}} \left\| f_1(\mathbf{x}_i) - \sigma(k \cdot f_2(\mathbf{x}_i)) \right\|^2\end{aligned}\quad (5-4)$$

5.2.2 基于双任务一致性的半监督学习

在半监督学习中, 首先定义 $\mathcal{D}_l$ 和 $\mathcal{D}_u$ 分别是有标签和无标签的数据集, 且 $\mathcal{D}_u \gg \mathcal{D}_l$ 。其中 $\mathcal{D} = \mathcal{D}_l \cup \mathcal{D}_u$ 为整个数据集。将有标签数据对表示为 $(\mathbf{X}, \mathbf{Y}) \in \mathcal{D}_l$, 则无标签数据表示为 $\mathbf{X} \in \mathcal{D}_u$, 其中 $\mathbf{Y}$ 是分割金标准。将三维体素及其类别对表示为 $(x, y) \in (\mathbf{X}, \mathbf{Y})$ 。对于有标签数据 $\mathcal{D}_l$, 分割任务的监督类型通常为全监督, 损失函数通常为 Dice, 定义如下:

$$\begin{aligned}\mathcal{L}_{Seg}(\mathbf{x}, \mathbf{y}) &= \sum_{\mathbf{x}_i, \mathbf{y}_i \in \mathcal{D}_l} \mathcal{L}_{Dice}(\mathbf{x}_i, \mathbf{y}_i) \\ &= \sum_{\mathbf{x}_i, \mathbf{y}_i \in \mathcal{D}_l} \left(1 - \frac{2 \sum_{x_j \in \mathbf{x}_i, y_j \in \mathbf{y}_i} f_1(\mathbf{x}_i) y_i}{\sum_{x_j \in \mathbf{x}_i, y_j \in \mathbf{y}_i} f_1(\mathbf{x}_j) + \sum_{y_j \in \mathbf{y}_i} y_j} \right)\end{aligned}\quad (5-5)$$

其中 $\sum_{x_j \in \mathbf{x}_i, y_j \in \mathbf{y}_i}$ 表示在三维图像中的体素求和, $\sum_{\mathbf{x}_i, \mathbf{y}_i \in \mathcal{D}_l}$ 表示在数据集中所有图像求和。然后定义了水平集表示回归任务的损失函数, 即预测水平集 $f_2(\mathbf{x})$ 和分割金标准转换而来的水平集金标准 $\mathcal{T}(\mathbf{y})$ 之间的 $\mathcal{L}_2$ 损失:

$$\mathcal{L}_{LSF}(\mathbf{x}, \mathbf{y}) = \sum_{\mathbf{x}_i, \mathbf{y}_i \in \mathcal{D}_l} \left\| f_2(\mathbf{x}_i) - \mathcal{T}(\mathbf{y}_i) \right\|^2 \quad (5-6)$$

算法 5-1 基于双任务一致性的半监督医学图像分割方法训练过程

Data: 来自不同数据集的标注图像及其金标准 $(\mathbf{x}_i, \mathbf{y}_i) \in \mathcal{D}_l$ 和未标注图像 $\mathbf{x}_i \in \mathcal{D}_u$, 双任务模型的分割分支参数 θ_1 , 回归分支参数 θ_2 以及它们共享的基础模型 θ , 其中 $f_1(\mathbf{x}) = \theta + \theta_1, f_2(\mathbf{x}) = \theta + \theta_2$

Result: 训练得到的图像分割模型

- 1 开始训练;
- 2 **while** 模型未收敛或训练未停止 **do**
- 3 采样最小批次数据 $b_l = (\mathbf{x}_i, \mathbf{y}_i) \in \mathcal{D}_l$ 和 $b = b_l + b_u$, 其中 $b_u = \mathbf{x}_i \in \mathcal{D}_u$
- 4 根据公式5-1, 生成水平集表示金标准 $\mathcal{T}(\mathbf{y}_i)$
- 5 生成两个任务各自的预测结果 $f_1(\mathbf{x}_i)$ and $f_2(\mathbf{x}_i), i \in \{1, \dots, N\}$ 其中 N 表示最小批次大小
- 6 根据公式5-2生成 $\mathcal{T}^{-1}(f_2(\mathbf{x}_i))$
- 7 计算 $\mathcal{L}_{DTC}(\mathbf{x}) = \frac{1}{|b|} \sum_{\mathbf{x}_i \in b} \|f_1(\mathbf{x}_i) - \mathcal{T}^{-1}(f_2(\mathbf{x}_i))\|^2$
- 8 计算 $\mathcal{L}_{LSF}(\mathbf{x}, \mathbf{y}) = \frac{1}{|b_l|} \sum_{\mathbf{x}_i, \mathbf{y}_i \in b_l} \|f_2(\mathbf{x}_i) - \mathcal{T}(\mathbf{y}_i)\|^2$
- 9 计算 $\mathcal{L}_{Seg}(\mathbf{x}, \mathbf{y}) = 1 - \frac{1}{|b_l|} \sum_{\mathbf{x}_i, \mathbf{y}_i \in b_l} 2 \frac{\sum f_1(\mathbf{x}_i) \mathbf{y}_i}{\sum f_1(\mathbf{x}_i) + \sum \mathbf{y}_i}$
- 10 计算 $\mathcal{L}_{total} = \mathcal{L}_{Seg} + \mathcal{L}_{LSF} + \lambda_d \mathcal{L}_{DTC}$
- 11 计算 $\mathcal{L}_{total}$ 的梯度, 然后根据梯度回传更新 θ_1, θ_2 和 θ 的参数.
- 12 **end**

值得注意的是, 对于带标注的图像, 水平集回归任务的金标准可以通过上述任务转换函数 $\mathcal{T}$ 从分割标注中 $\mathbf{Y}$ 自动生成。最终的优化函数定义如下:

$$\mathcal{L}_{total} = \mathcal{L}_{Seg} + \mathcal{L}_{LSF} + \lambda_d \mathcal{L}_{DTC} \quad (5-7)$$

其中, $\mathcal{L}_{Seg}$ 和 $\mathcal{L}_{LSF}$ 仅用于有标签数据, $\mathcal{L}_{DTC}$ 在训练过程中同时作用于有标签和无标签数据, 因此两个任务可以在半监督的方式下使用有标签数据或无标签数据来联合优化网络参数。根据既往工作^[94,97], 该方法使用与时间相关的高斯预热函数 $\lambda_d(t) = e^{(-5(1-\frac{t}{t_{max}})^2)}$ 来控制监督损失和无监督一致性损失之间的平衡, 其中 t 表示当前的迭代次数, t_{max} 是总的迭代次数。基于双任务一致性的半监督医学图像分割算法训练流程如算法5-1所示。

5.3 实验验证

5.3.1 实验数据及实施细节

为了验证本章所提出方法的有效性, 本工作在广泛使用的两个公开医学图像半监督分割基准数据集上进行了大量实验, 并与目前几种最先进的医学图像半监督分割方法就行了性能与开销比较。

1) 实验数据集: 第一个数据集是左心房 MRI 数据集^[169], 其中包含 100 个 3D 钆增强 MRI 图像, 分辨率为 $0.625\text{ mm} \times 0.625\text{ mm} \times 0.625\text{ mm}$ 。本实验沿用了以外工作的标准实验设置^[97,171], 将其中 80 个数据用于训练, 剩下的 20 个数据用于验证, 并采用了相同的预处理方法进行数据处理。第二个数据集是腹部胰腺 CT 数据集^[172], 其中包含了 82 个腹部 CT 图像。遵循 Xia 等人^[157] 的既往工作, 本实验将该数据集划分为 62 个图像用于训练, 20 个图像用于测试。在预处理中, 本文使用软组织 CT 窗宽窗位来确定图像灰度范围, 并将所有图像重新采样为各向同性分辨率为 $1.0\text{ mm} \times 1.0\text{ mm} \times 1.0\text{ mm}$ 。最后, 基于标注金标准裁剪图像成以胰腺为中心区域往外扩展 25 个体素的补丁, 然后并将它们标准化成零均值和单位方差的灰度分布。在这项工作中, 报告了所有方法在 10% 或 20% 有标签图像和 90% 或 80% 无标签图像的实验设置下的分割性能, 这些都是典型的半监督学习实验设置^[97,157,171]。

2) 实施细节: 本研究所有实验均在一个带有四张 NVIDIA 1080TI GPUs 的 Ubuntu20.04 服务器上通过 PyTorch1.8^[189] 实现和完成。本实验遵循基准实验设置, 使用 VNet^[5] 作为所有实验的骨干网络, 并通过在原始 VNet 的末端添加一个新的回归层来完成双任务预测。所有模型均由 SGD 优化器训练了 6000 次得到, 其中初始学习率 (lr) 为 0.01, 每 2500 次迭代衰减 0.1。最小批次大小 (mini-batch) 为 4, 包含 2 个有标签图像和 2 个无标签图像。受到 Xu 等人^[209] 工作的启发, 本工作将 k 值设置为 1500。本实验随机裁剪大小为 $112 \times 112 \times 80$ 的补丁 (左心房分割) 和大小为 $96 \times 96 \times 96$ 的补丁 (腹部胰腺 CT) 作为网络输入。为了避免过拟合, 在训练阶段使用标准的在线数据增强方法进行数据增强^[97]。需要注意的是, 在本实验中, 水平集函数在训练阶段之前生成, 而不是即时生成, 因为水平集函数具有变换不变的, 这样可以降低计算开销, 加速训练过程。在推断阶段, 使用滑动窗口策略获得最终结果, 对于左心房数据步长为 $18 \times 18 \times 4$, 对于腹部胰腺数据步长为 $16 \times 16 \times 16$ 。在推理阶段, 本实验使用了像素级分类分支的输出作为分割结果。为了公平比较, 本实验不使用任何后处理或集成方法。遵循已有工作的实验设置^[97], 本实验使用四个指标定量评估所有方法的分割结果, 包括 Dice 系数、Jaccard 系数、平均表面距离 (ASD) 和 95% 豪斯多夫距离 (95HD)。

5.3.2 算法性能比较

在本章中, 将提出的方法与六种最先进的医学图像半监督分割方法进行了全面的比较, 包括深度对抗网络 (DAN)^[100]、熵最小化方法 (Entropy Mini)^[204]、交叉一致性训练方法 (CCT)^[158]、均值教师自我集成模型 (MT)^[94]、不确定性感

知均值教师模型 (UA-MT) [97] 和形状感知对抗网络 (SASSNet) [171]。值得注意的是, 本实验使用了 DAN、MT、UA-MT 和 SASSNet 的官方代码和实验结果, 并重新实现了 Entropy Mini 和 CCT 用于医学图像分割, 由于 GPU 内存的限制, CCT 的实现中使用一个主解码器和三个辅助解码器。

首先在腹部胰腺 CT 数据集上评估所提出的方法。表 5-1 展示了这些方法的定量比较结果。与仅使用 12 个有标签图像训练的全监督 VNet 相比, 所有利用未标记图像的半监督方法都显著改善了分割性能。MT、UA-MT 和 CCT 的性能略优于 Entropy Mini 和 DAN, 表明基于扰动的一致性损失对半监督分割问题是有帮助的。此外, UA-MT 比 MT 更好, 因为不确定性修正伪标签可以有效地指导学生模型的学习过程进而实现更好的分割结果。SASSNet 在现有方法中实现了最佳性能, 表明形状先验有利于提升医学图像半监督分割的效果。值得注意的是, 本章所提出的方法在所有评价指标上均取得了比最先进的半监督方法更好的实验性能, 而无需使用复杂的多网络架构, 这证实了双任务一致性具有较好的建模能力进而从未标记的数据中提取丰富的信息。与此同时, 该方法不需要依赖任何多重推理或迭代更新方案, 进一步降低了计算内存成本和运行时间。

此外, 本实验进一步在公开左心房 MRI 数据集上验证了所提出的方法, 这是一个广泛用于半监督医学图像分割方法验证的数据集 [97, 171]。这些方法的定量比较如表 5-2 所示。可以发现, 该方法在所有评估指标上都比其它方法取得了最佳的准确性, 特别是在 ASD 和 95HD 方面。具体来说, 当使用 8 和 16 个标注数据时, 与 VNet 下限相比, 该方法的 Dice 结果分别提升了 7.52% 和 3.39%。此外还可以发现该方法利用 16 个标注数据取得的实验结果与全监督 VNet 利用 80 个标注数据的结果已经非常接近, 说明了提出方法的有效性。图 5-2 展示了一些胰腺分割和左心房分割的可视化结果。与其它方法相比, 本章方法产生的分割结果与金标准有更高的重叠率, 且假阳性更少并保留了器官更多的解剖学细节, 这进一步表明了提出方法的有效性、泛化性和鲁棒性。此外, 本实验还研究了不同方法的训练成本。网络参数和训练时间的定量比较如表 5-1 和表 5-2 所列。可以观察到, 本方法所需的训练时间比 MT、DAN、UAMT、CCT 和 SASSNet 少, 因为本方法使用了一个更简单的分割网络, 参数较少, 并且不需要在迭代中多次推理同一图像。与 Entropy Mini 和完全监督基线相比, 本方法在具有可比的计算成本的情况下取得了更好的准确性。因此, 本实验证明了本章方法在准确性、网络参数和计算成本之间达到了最佳的平衡。

表 5-1 在腹部胰腺 CT 数据集上，本章提出方法与其它医学图像半监督分割方法的定量比较。第一行和第二行是本实验的全监督学习结果的下限和上限，最后一行是本章提出的方法，其它是以前的方法

	标注数据	未标注数据	Dice (%)	Jaccard (%)	ASD (voxel)	95HD (voxel)	模型参数 (M)	训练时间 (h)
VNet	12	0	70.63	56.72	6.29	22.54	9.44	2.1
VNet	62	0	81.78	69.65	1.34	5.13	9.44	2.3
MT (NeurIPS'17)	12	50	75.85	61.98	3.40	12.59	9.44	2.9
DAN (MICCAI'17)	12	50	76.74	63.29	2.97	11.13	12.09	3.3
Entropy Mini (CVPR'19)	12	50	75.31	61.73	3.88	11.72	9.44	2.2
UA-MT (MICCAI'19)	12	50	77.26	63.82	3.06	11.90	9.44	3.9
CCT (CVPR'20)	12	50	76.58	62.76	3.69	12.92	15.65	4.1
SASSNet (MICCAI'20)	12	50	77.66	64.08	3.05	10.93	20.46	3.9
Ours	12	50	78.27	64.75	2.25	8.36	9.44	2.5

表 5-2 左心房 MRI 数据集上，本章方法与其它医学图像半监督分割方法在不同标注比例下的定量比较。第一，二行和第三行是本实验的全监督学习结果的下限和上限，最后一行是本章提出的方法，其它是以前的方法

	标注数据	未标注数据	Dice (%)	Jaccard (%)	ASD (voxel)	95HD (voxel)	模型参数 (M)	训练时间 (h)
VNet	8	0	79.99	68.12	5.48	21.11	-	-
VNet	16	0	86.03	73.26	5.75	17.93	9.44	1.8
VNet	80	0	91.14	83.32	1.52	5.75	9.44	2.0
UAMT(MICCAI'19)	8	72	84.25	73.48	3.36	13.84	-	-
SASSNet(MICCAI'21)	8	72	87.32	77.72	2.55	9.62	-	-
LG-ER-MT(MICCAI'20)	8	72	85.54	75.12	3.77	13.29	-	-
DUWM(MICCAI'20)	8	72	85.91	75.75	3.31	12.67	-	-
Ours	8	72	87.51	78.17	2.36	8.23	-	-
MT(NeurIPS'17)	16	64	88.23	79.29	2.73	10.64	9.44	3.2
DAN (MICCAI'17)	16	64	87.52	78.29	2.42	9.01	12.09	3.7
Entropy Mini (CVPR'19)	16	64	88.45	79.51	3.72	14.14	9.44	1.9
UA-MT (MICCAI'19)	16	64	88.88	80.21	2.26	7.32	9.44	3.6
LG-ER-MT (MICCAI'20)	16	64	89.62	81.31	2.06	7.16	-	-
DUWN (MICCAI'20)	16	64	89.65	81.35	2.03	7.04	-	-
CCT (CVPR'20)	16	64	88.83	80.06	2.49	8.44	15.65	3.9
SASSNet (MICCAI'20)	16	64	89.27	80.82	3.13	8.83	20.46	4.4
Ours	16	64	89.42	80.98	2.10	7.32	9.44	2.2

表 5-3 本章方法的不同组成部分在腹部胰腺 CT 数据集上的消融实验

	标注数据	未标注数据	Dice (%)	Jaccard (%)	ASD (voxel)	95HD (voxel)	模型参数 (M)	训练时间 (h)
Seg	12	0	70.63	56.72	6.29	22.54	9.44	2.1
LSF	12	0	71.78	57.55	6.31	20.74	9.44	2.1
Seg + LSF	12	0	73.08	58.65	4.47	18.04	9.44	2.2
Seg + LSF + DTC	12	0	74.84	60.78	2.17	9.34	9.44	2.3
Seg	62	0	81.78	69.65	1.34	5.13	9.44	2.3
LSF	62	0	82.25	70.23	1.18	5.19	9.44	2.5
Seg + LSF	62	0	82.46	70.61	1.22	4.97	9.44	2.5
Seg + LSF + DTC	62	0	82.80	71.05	1.45	4.67	9.44	2.5

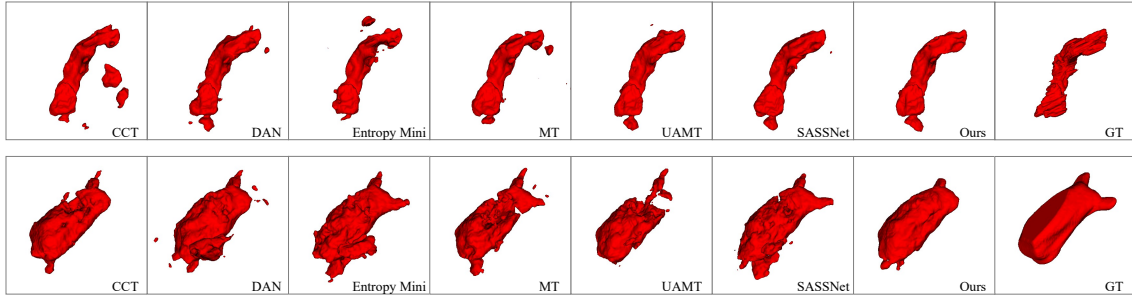


图 5-2 不同医学图像半监督分割方法在 20% 标注数据下模型预测结果的 3D 可视化比较。第一行是腹部胰腺分割结果，第二行是左心房分割结果

5.3.3 消融实验

为了确定提出方法各个组成部分的有效性，本消融实验首先调研了两种不同任务单独学习时各自的性能表现和进一步双任务联合学习时的实验表现，以及本章提出的双任务一致性策略的分割表现。该实验在腹部胰腺 CT 数据集上分别使用了 12 个有标签数据和 62 个有标签数据来训练分割网络，并比较了四种不同的模型训练策略：1) 仅考虑任务 1，即分割分支 (Seg)，2) 仅考虑任务 2，即回归分支 (LSF)，3) 使用双分支网络同时考虑任务 1 和任务 2 (Seg+LSF)，以及 4) 本章提出的双任务一致性方法 (Seg+LSF+DTC)。

表5-3展示了这些方法在不同标注数据量情况下的分割结果。从表中可以发现，在两种不同标注量的实验设定下，水平集回归任务均能比分割任务取得更好的实验结果，分割任务和回归任务的 Dice 分别为 70.63%，71.78% 和 81.78%，82.25%。从表中还可以观察到在距离相关的指标方面，水平集回归任务比分割任务能获得更好的实验结果，尤其是在标注数据不足的情况下，回归任务在 ASD 方面减少了 2.3 个体素在 95HD 方面减少了 8.7 个体素。此外，两个任务联合学习取得了也比它们各自单独学习更好的实验结果，尤其是在 12 个有标签数据的实验设定下，联合学习比它们单独学习 Dice 分别提高了 2.45% 和 1.30%。进一步发现，本章提出的双任务一致性可以在联合学习的基础上继续提高模型的分割性能，尤其是在标注数据较少的情况下 Dice 比联合学习提高了 1.72%，ASD 和 95HD 也取得了略好或者是差不多的分割性能。

图 5-3展示了四种不同训练策略的三维可视化结果。从这张图可以看出，水平集回归任务能比分割任务取得更少的假阳预测，并保存更好的器官解剖结构。此外，回归任务和分割任务的联合学习能促进模型的学习过程，获得更好的分割结果。进一步，本章提出的双任务一致性在减少了过分割的同时更好的保持了器官的形状特点，与金标准的相似度更好。这些结果也说明了从可视化的角度来说，

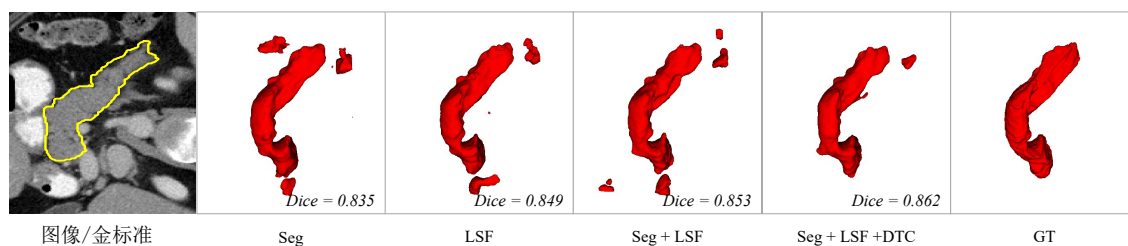


图 5-3 腹部胰腺分割数据集在不同训练方法情况下预测结果的 3D 可视化比较。该实验仅使用 12 个有标签图像进行训练，没有使用未标记的图像。

Seg: 单一分割任务的预测结果, LSF: 单一水平集表示回归结果, Seg+LSF: 双任务网络中分割分支的预测结果, Seg+LSF+DTC: 基于双任务一致性的双任务网络中分割分支的预测结果, GT: 金标准

可视化结果和数值定量分析有着相同的趋势，进一步显示了本文提出的双任务一致性的优越性。

本实验进一步在腹部胰腺 CT 数据集进行了关于数据利用效率的研究，并将本章将提出的双任务一致性半监督分割方法与仅在标注数据上全监督学习和在标注数据上引入双任务一致性的全监督学习进行了比较。图5-4展示不同标注数据比例下模型取得分割结果的 Dice 分数折线图。可以观察到，在不同标注数据的实验设置下，半监督方法始终比全监督方法表现更好，表明本章双任务一致性方法有效利用了未标记的数据并带来了性能提升。同时，本实验还发现随着标注数据规模的提升，全监督方法与半监督方法之间的性能差距逐渐缩小，和主流的研究结果趋势一样。此外，当标注数据规模较小时，本章提出的方法也可以获得比全监督方法更好的分割结果，表明了本章提出的方法能有效减少模型对标注数据规模的依赖，并进一步降低模型开发的标注成本，有潜力为临床应用开发提供一种低标注成本的分割算法。

5.4 讨论与分析

从医学图像中准确且鲁棒地分割器官或病变区域，在诊断和治疗规划等许多临床应用中发挥着重要作用。随着标注数据的不断增长，深度学习在图像自动分割方面取得了长足的进步。然而，对于医学图像来说，由于需要专业知识和时间来进行图像标注，因此三维医学图像分割标注往往非常的昂贵。为了降低标注成本，本章提出了一种利用较少的标注数据来开发高性能的医学图像分割模型的算法，即通过结合有限数量的标注数据和大量的未标注数据进行学习，进而获得准确的分割结果。具体而言，本文提出了一种在任务层面构建一致性的正则化方法，

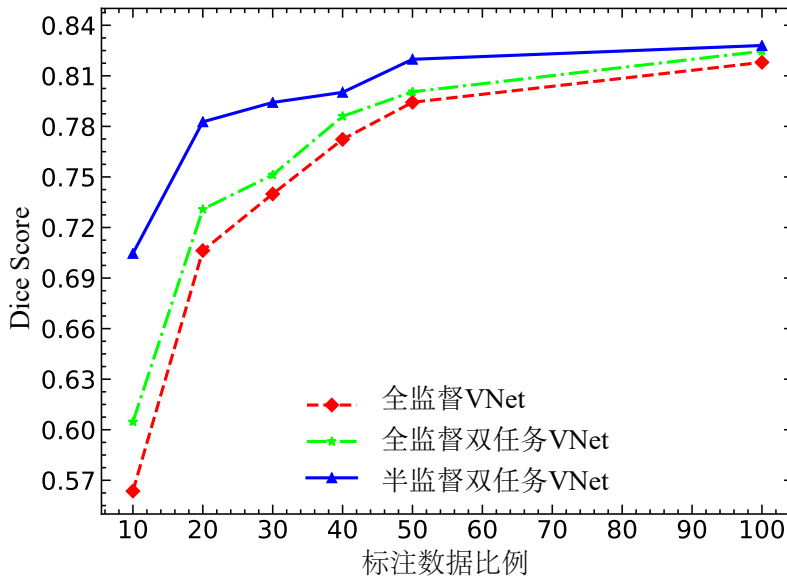


图 5-4 双任务一致性方法在不同标注数据比例下的腹部胰腺分割性能曲线图。红色虚线和绿色虚线分别展示了全监督 VNet 和双任务 VNet 的分割结果，它们仅使用了标记数据进行训练。蓝色实现展示了基于双任务一致性的半监督分割结果

通过将不同任务的不同预测结果对齐到共同的表示空间，然后鼓励它们在共同表示空间内保持一致。实现上述目标的最主要挑战是如何以可微的形式进行不同任务的表示转换，即可以在模型训练过程中构建一致性损失，并可以梯度回传进行模型训练。为此，本文首次将传统分割表示方法水平集引入到半监督学习中，并构造了基于分割和水平集回归的双任务一致性。在半监督医学图像分割任务中，该方法可以有效的促进模型从无限标注数据中进行学习，提高模型的分割性能的同时减少数据标注成本。

与以往基于数据一致性的半监督医学图像分割方法不同，它们通过向未标记图像添加微小扰动，然后强制模型的原始数据和扰动数据预测保持一致，本文提出了一种新的任务层面的一致性构造策略来进行半监督学习。该方法较以往方法有两大优势：1) 该方法不需要设计数据层面的一致性，不依赖特定扰动和模型；2) 该方法显式的建模了多任务之间的表示差异，有利于融合不同任务之间的表示差异，例如水平集考虑了更多分割目标的全局解剖和几何特征。任务层面，同一任务可以被视做不同的表示，例如分割任务可以同时被看做像素层面的分类，用分类结果表示分割结果；也可以用水平集回归任务表示分割任务，用零水平内的区域表示分割结果。因此，两个任务之间本身就存在表示差异，可以直接被视为任务层面的扰动。此外，基于像素分类的分割表示只考虑每个像素单独的分类结

果没有显式的建模全局信息，相反水平集表示基于全局带符号的距离来显式建模分割对象，因此两个任务可以在分割表示方面互补。

两个公开半监督分割基准上的实验结果也证明了本方法的有效性。首先在左心房分割 MRI 数据集上，本方法在 8 和 16 个标注数据两个广泛使用的实验设置下均优于现有方法，如表5-2所示。与此同时，在腹部胰腺 CT 数据集上，本方法也优于现有方法，如表5-1所示。同时消融实验结果也详细的给出了每个组成部分的具体贡献，进一步证明了本方法的有效性。此外，在模型参数和计算开销方面，与以往方法相比本方法在训练阶段并没有显著增加模型的训练成本，在测试阶段本方法没有增加任何额外的计算开销，证明了该方法的实用性。

尽管本方法在两个不同模态和器官的数据集上证明了其有效性，但其也存在一些局限性。首先是，本方法中的水平集回归需要显式建模分割对象的解剖结构，因此其更适合用于解剖结构固定的正常器官，对解剖结构不固定的病灶实验结果缺乏验证。未来本研究将继续探索该策略在不同分割场景下的适用性。其次是本方法受限于不同任务之间的可微变换，但直接构建不同任务之间的可微变换非常困难，需要保证其存在可微的数学表达。作为替代，本章计划探索用深度学习模型来代替传统的数学可微性保证，即利用模型学习构建不同任务之间的可微转化，例如利用一个深度学习在有标注数据上训练模型直接将的分割结果转化从水平集(该过程不可微)或水平集转换成分割结果。本方法只在二类分割任务上进行了验证，缺乏对多类分割任务的验证，但其可以通过简单修改扩展到多类分割任务中。在多类分割任务中，分割分支不需要做任何调整；只需要将水平集回归分支扩展为多个单类别回归分支，即使用类别数目个通道来回归水平集表示，并按照类别对应通道数与分割分支结果进行一致性约束即可。此外，该方法只在三维图像上进行了实验验证，未来探索该方法在二维图像上的有效性也有利于进一步推广该方法进而减少模型开发的数据标注成本。

5.5 本章总结

本章提出了一种简单有效的基于任务级一致性的半监督医学图像分割方法，其通过双任务一致性实现。本章使用双分支网络，同时预测像素级分类结果和目标对象的水平集表示，该表示能够捕获全局层面的形状和几何信息。为了构建半监督训练框架，本章通过任务变换层在分类预测和水平集回归预测之间构建双任务一致性。该方法在两个三维医学图像数据集上实现了最先进的结果，包括左心房分割 MRI 数据集和腹部胰腺 CT 数据集。卓越的性能证明了提出方法的有效性、稳健性和泛化性。目前，这项工作专注于单类别分割以简化实验验证。然而，该方

法可以直接扩展到多类别情况。此外，该方法还可以轻松地扩展到使用额外的任务，如边缘提取^[210]和关键点估计^[211]，只要两个任务之间存在可微的变换。此外该方法还可以引起计算机视觉社区对不同任务一致性的关注，因为在许多方向上都可以构造任务一致性以半监督学习的方式从大量未标记数据中学习，例如两流视频识别^[212]、多任务图像重建^[213,214]、图像分割和边缘检测等。在未来，本研究将继续把该方法扩展到更多的计算机视觉任务中，以减少标注成本，并进一步研究融合策略，将所有不同任务的预测结果集成为更好的性能。

第六章 基于主动学习的跨域医学图像分割及临床验证

6.1 引言

以上三章主要是探索如何利用已有的有限标注数据开发高性能医学图像分割模型，大量实验证明了这些方法可以有效利用这些有限的标注数据训练得到性能良好的分割模型。上述工作的良好表现主要依赖于以下两个前提，首先是实验的训练和测试集中所有数据分布一致，其次是所有数据的标注标准一致。但在实际临床场景中，尤其是多中心推广过程中，以上两个假设很难满足，因此探索如何用有限的标注数据来实现跨中心和标注者的准确分割变得极为重要。此外在实际应用中，根据特定医院和标注需求进行少量数据标注然后进行模型微调可以显著提高模型在特定场景下的性能。根据临床需求和实际应用挑战，针对此问题，本章从问题设计的方向着手探索如何利用在其他中心数据上训练好的分割模型和目标域上有限的标注数据，通过模型微调来实现多中心高性能泛化。目前，无源域适应被广泛用于减少数据标注成本和提高模型在未知数据分布中心上的泛化性 (Source-Free Domain Adaption, SFDA)，即利用源域上训练的模型去未知目标域上进行无监督模型微调然后用于测试。然而这种学习策略在实际场景中表现受限，尤其是在真实的临床场景下该策略取得的结果与临床需要的结果差距较大。此外，迁移学习通过将源域数据和目标域数据的少量标注数据进行联合学习也可以提高模型的迁移性能。但相比于无源域适应，多中心迁移学习在每个数据中心上都需要获取源域数据和标注，这种设置不能保护源域数据的隐私。

针对上述问题，本章首先从临床需求的角度出发提出了一种新的同时考虑数据隐私保护和降低数据标注成本的学习场景，无源主动域适应 (Source-Free Active Domain Adaption, SFADA)，即结合源域训练的模型和主动学习从目标域选择少量具有代表性的样本进行标注，然后结合半监督学习对预训练模型进行微调以提高模型泛化性。不同于无源域适应和多中心迁移学习，该场景同时考虑了二者的优点，在不泄露源域数据隐私的情况下有效的减少了泛化需要的标注成本。图 6-1 展示了传统多中心迁移学习与本章提出的无源主动域适应学习的训练差异。从图中可以发现，要实现基于有限标注数据的无源主动域适应存在两个主要挑战，首先是如何只利用源域训练好的模型从目标域的数据中选出信息量最丰富的样本进行标注；其次是如何利用这些少量带标注的高信息量数据进行模型微调。

针对这两大挑战，本章首先提出了一种同时考虑源域和目标域差异的双参考主动样本选择策略来同时选择目标域无关样本和目标域相关样本进行标注

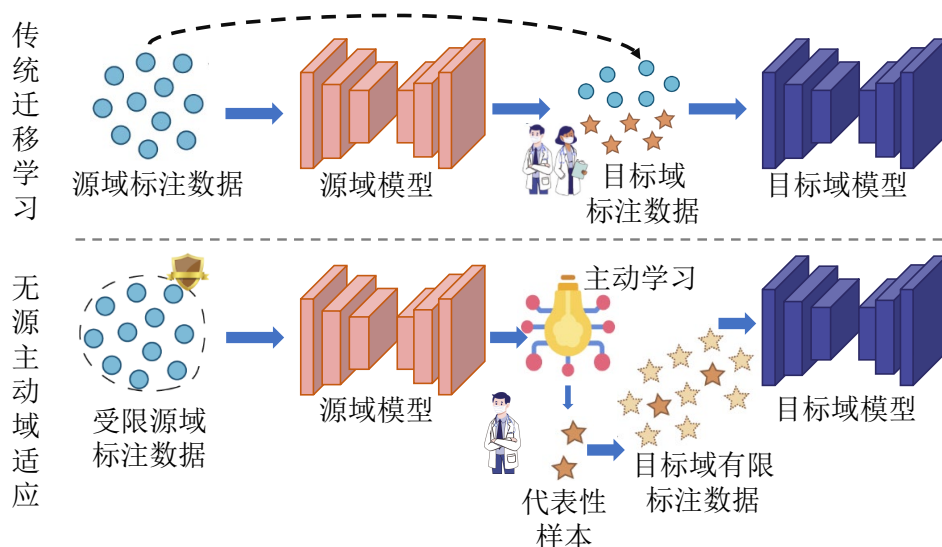


图 6-1 传统多中心迁移学习与本章提出的无源主动域适应（SFADA）学习的流程差异。本方法旨在减少图像标注的同时避免源域数据泄露

（Source-Target-domain Dual-Reference, STDR）。传统主动学习要求所有数据必须服从统一分布，而本章考虑了更有挑战性的问题，即从不同分布的数据中选择高信息量样本。因为真实临床场景下，不同中心的数据往往分布存在较大差异。同时，本章提出了一种多阶段半监督学习策略来结合预训练模型和带标注的主动样本来进行模型微调。不同于传统半监督学习方法，多阶段半监督学习可以利用每个阶段更新后的数据进行多次模型微调，有利于减少数据分布差异带来的性能降低。

为了证实该方法的有效性，本工作构建了两个数据集来进行实验验证，包括来自五个中心 1057 名鼻咽癌患者的多中心鼻咽癌靶区勾画数据集（Gross Tumor Volume, GTV），和来自 170 名鼻咽癌患者和四个专家标注的多标注者数据集。这两个数据集分别在图像和标注层面存在着数据分布差异。大量实验结果表明，与最先进的域适应方法（无论是否访问源数据）以及现有主动学习方法相比本章提出的方法均表现最佳。本工作的主要贡献可概括如下：

- 1) 本章首次提出了无源主动域适应的概念，为基于有限标注的多中心泛化提供了一种新的可行的解决方案。
- 2) 本章提出了一种基于目标域无关和目标域相关双参考的样本选择方法，用于从目标域中主动选择代表性样本进行标注，并进一步降低标注成本，而且无需依赖源域数据。
- 3) 本章采用了多阶段半监督学习方法，利用通过双参考策略选择的标注样本和剩余无标注样本进行共同训练模型。使模型能够从更多样化的数据中学习，

从而提高了分割的准确性。

- 4) 在多中心和多标注者鼻咽癌靶区勾画数据集上的大量实验结果表明，本章提出的方法利用有限标注数据很好地适应了不同数据分布和不同勾画风格的分割任务。此外，多中心临床验证进一步证明了该方法只需要少量的医生介入即可达到临床可用的结果。

6.2 算法设计

6.2.1 基于目标域无关和目标域相关的双参考主动样本选择

在域适应、域泛化任务中，根据数据分布、标注分布等差异将数据分为源域（source）和目标域（target）。其中，源域和目标域对应的图像和标注分别被表示为 (F, Y^s) 和 (F', Y') ，在源域数据上训练的模型被称为源域模型 (θ_s) ，在目标域数据上微调的模型被称为目标域模型 (θ_t) 。通常，在无源域适应条件下，源域数据只能用于模型开发，模型开发好后源域数据即不可被获取，此外目标域的数据没有标注，只能利用源域模型进行微调等。本章在此基础上，引入了新的主动标签选择和标注策略，从目标域上选取少量样本进行标注，进而提高模型在目标域上的表现。

图 6-2 呈现了本章提出 STDR 框架的示意图。在跨中心图像分割中，一个重要的挑战是不同中心之间数据分布存在域差异，同时仍然共享一些常见的解剖结构知识。然而，仅在从目标数据中选择少量相似的主动样本进行学习很容易导致过拟合 [215]。因此，本章提的 STDR 策略通过联合考虑源域和目标域数据来选择源域代表性样本 P^s 和目标域代表性样本 P' 。这种方法具有两个关键优势：在目标域微调期间，样本 P^s 在不访问源数据的情况下增强了共享的语义知识，并有助于学习目标域无关的表示；同时，样本 P' 有助于将模型有效迁移到目标域数据，从而克服潜在的域差异挑战。图 6-3 展示了 STDR 选择的一些 MRI 图像示例。

在无源域适应情景下，本章首先通过全监督的方式在源域图像 F 及其对应的金标准 Y^s 上训练分割网络。其中通过最小化损失函数更新网络参数 θ_s ，本章采用了一个 Dice 和交叉熵结合的联合损失函数 Ψ ，其定义如下：

$$\Psi(y_s(i), p_s(i, \theta)) = - \sum_{i=1}^N y_s(i) \log p_s(i, \theta) + 1 - \frac{2 \sum_{i=1}^N y_s(i) \hat{y}_s(i)}{\sum_{i=1}^N y_s(i)^2 + \sum_{i=1}^N \hat{y}_s(i)^2}, \quad (6-1)$$

其中，网络在源域图像 F 像素 i 处的输出被表示为 $p_s(i, \theta) \in [0, 1]$ ， N 表示每张图像所有像素的总数。

- 1) 样本相似性排序：**考虑到在医疗中心中存在多位医生进行数据标注，并且

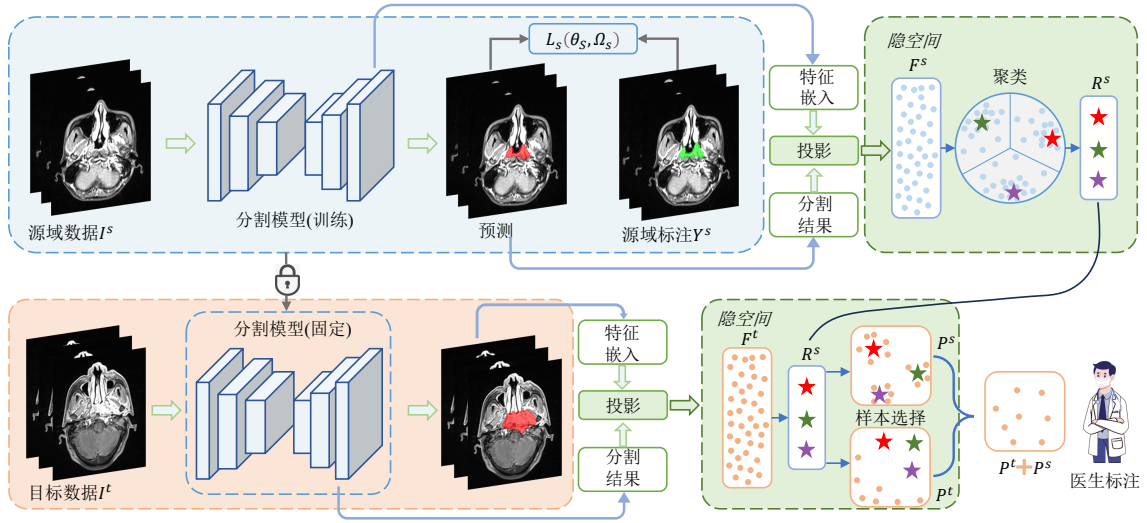


图 6-2 本章提取主动样本选择策略 STDR 的流程概述。给定源域图像 F 以及标注 Y ，STDR 首先在源域数据中训练分割网络。然后将分割特征映射到隐空间，得到 F^s ，聚类算法聚类出参考点 R^s 。目标域中的数据经过类似的操作，得到 F^t 。然后选择两种类型的代表性样本，由放射肿瘤学家手动标记

可能数据可能来自不同供应商的各种成像设备，因此有必要选择多个参考点 R^s 来有效表示这些数据和标注层面的分布特征。这种方法确保了数据分布的全面表示，并考虑了不同医生和成像设备引入的潜在变化。为了实现这一点，本章首先冻结上述预训练的分割网络的参数。然后，将网络的倒数第二层的特征嵌入与相对应的预测进行合并，然后共同映射到一个隐空间。具体过程如下：

$$F^s(F(i)) = \frac{1}{\|N^s\|} \text{Max}P_f(\text{Avg}P_c(\hat{Y}^s(i) \otimes f_E(F(i)))), \quad (6-2)$$

其中， $\|N^s\|$ 表示属于分割目标的像素总数， $\hat{Y}_s(i)$ 表示第 i 个预测结果， $f_E(F(i)) \in R^{C \times H \times W}$ 表示第 i 个图像的倒数第二层输出特征嵌入。 $\text{Max}P_f$ 是对空间维度为 $R^{H \times W}$ 的图像特征进行最大池化操作，而 $\text{Avg}P_c$ 是沿着通道维度 R^C 进行的平均池化操作，将三维特征张量降维为二维。值得注意的是，在密集预测任务中，本文选择了一个较小的核大小进行最大池化，以从多个图像补丁中提取特征。与将整个图像池化为单个元素相比，这种方法提供了更全面的表示。最终的 F^s 是这个二维向量的展平结果（向量维度为 1×256 ）。

接下来，本文采用聚类方法（为了保证方法的通用性，本文采用最广泛应用的 K-Means 聚类^[216]）将所有源域图像的特征向量分组成 K 个不同的簇。这个过程对本章的方法非常重要，因为它有助于识别源域数据中的代表性特征，然后用

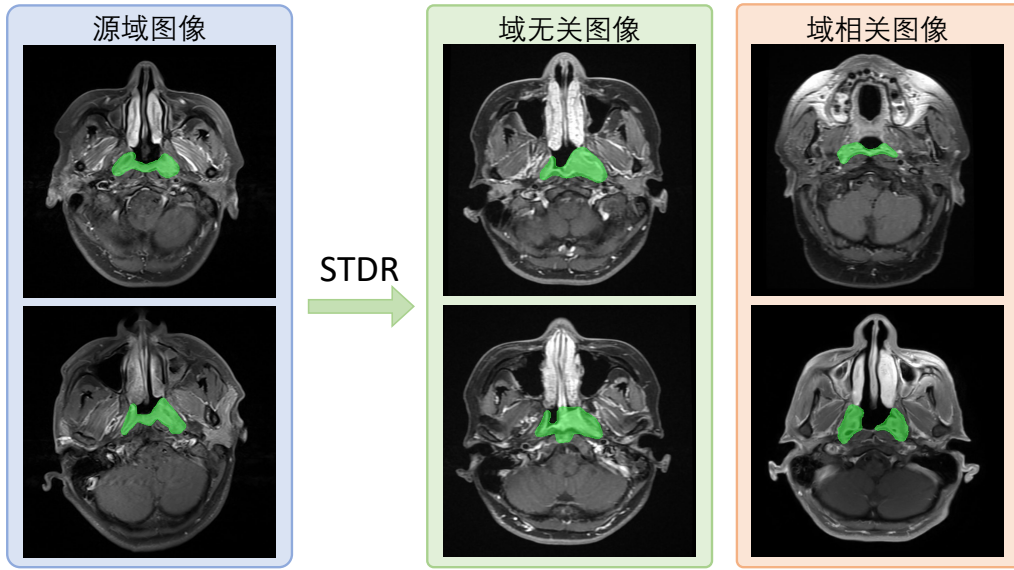


图 6-3 源域（SPH）和目标域（APH）中目标域无关和目标域相关样本可视化比较。源域图像与目标域无关图像相似性更高，而与目标域相关图像的差异更明显。STDR 为本文提出的源域目标域双参考选择策略

于进一步的分析和比较。聚类的目标是最小化以下误差：

$$\sum_{k=1}^K \sum_{i \in R_k} \|F^s(F(i)) - R_k^s\|_2^2, \quad (6-3)$$

其中， $\|\cdot\|_2^2$ 表示欧氏距离， R_k^s 表示簇 R_k 的代表性点，定义如下：

$$R_k^s = \frac{1}{\|R_k\|} \sum_{i \in R_k} F^s(F(i)). \quad (6-4)$$

其中， $\|R_k\|$ 表示属于簇 R_k 的图像数量。质心 R^s 在活动样本选择过程中起着重要的参考作用，用于比较和选择目标图像。

2) 目标域无关和目标域相关双参考主动样本选择：如图 6-2所示，源域中训练的模型参数被固定，并直接用于推理目标域中的图像。然后按照公式6-2中描述，将得到的特征映射到隐空间以获得 F^t 。接下来，计算 $F^t(F(i))$ 与所有源域参考 R^s 之间的欧氏距离，并将最小距离指定为目标域样本与源域相似性的度量（*Simi*）：

$$Simi(F^t(i)) = \min_k \|F^t(F(i)) - R_k^s\|_2^2, \quad (6-5)$$

其中， $F^t(\cdot)$ 类似于公式6-2，表示目标域图像的投影函数。根据源域特征和目标域图像的相似性排序可以将目标域图像分为与源域特征相似和不相似两类。进一步，本章将目标域图像特征与源域分布特征相似的样本视为目标域无关样本，因为他

们的特征没有随着域的改变而发生改变。同时，也将目标域图像特征与源域分布特征不相似的样本视为目标域相关样本，因为他们的特征随着域的改变而发生改变。

在本章的无源设置中，主动样本选择包括目标域代表性样本 P^t （目标域相关）和源域代表性样本 P^s （目标域无关）。这种方法确保了在迁移到目标域时保留了源域的共享目标域无关的语义知识。它们和目标域相关样本 $Sele(P^s, P^t)$ 组成如下所示：

$$Sele = Min(Simi)[\frac{p}{2}\%] + Max(Simi)[\frac{p}{2}\%], \quad (6-6)$$

其中 $Simi$ 表示所有目标域数据的相似性列表， $p\%$ 表示要选择用于主动学习的样本比例。随后的消融实验进一步证明了这种方法有效地保留了共享的目标域无关语义知识和引入了特定的目标域相关信息，从而提高了分割准确性。

6.2.2 基于多阶段半监督学习的无源主动域适应

本章在域适应过程中选择用三阶段的半监督学习来更新分割模型，如图6-4所示。第1阶段主要使用主动标记样本 I_{AL} 对模型参数进行完全监督微调，优化函数如下：

$$L_s(\theta_t, \Omega_t) = \frac{1}{|\Omega_t|} \sum_{i=1}^T \Psi(y_t(i), p_t(i, \theta_t)) \quad (6-7)$$

其中， θ_t 代表微调网络参数， $y_t(i) \in \{0, 1\}$ 表示来自 Y_{AL} 的像素标签， $p_t(i, \theta_t)$ 是网络的预测值。 Ψ 与公式6-1相同。随后，在第2阶段，保持这些微调的模型参数不变，并利用它们预测未标记数据 I_{UL} 的伪标签 Y_{PL} 。最后，在第3阶段，使用预测的伪标签 Y_{PL} 和主动标记样本 Y_{AL} 共同微调模型参数。值得注意的是，这里使用了与第1阶段相同的损失函数，只是它包含了更全面的目标域数据。上述多阶段过程使本方法能够有效地利用标记和未标记数据。通过对主动标记样本进行初始微调，模型能够学习目标域的特征分布，同时保留共享的目标域无关语义知识。这为模型奠定了坚实的基础，使其能够生成更高质量的伪标签。通过使用训练好的模型为未标记数据 I_{UL} 生成伪标签 Y_{AL} ，来扩展训练数据集，并获得了额外的有价值监督信息。在最后一个微调阶段中将预测的伪标签与主动选择的样本结合起来，有助于进一步改进模型，提高分割性能和多中心泛化能力。随后的消融实验将进一步验证这一方面。整个方法的训练过程如算法6-1所示。

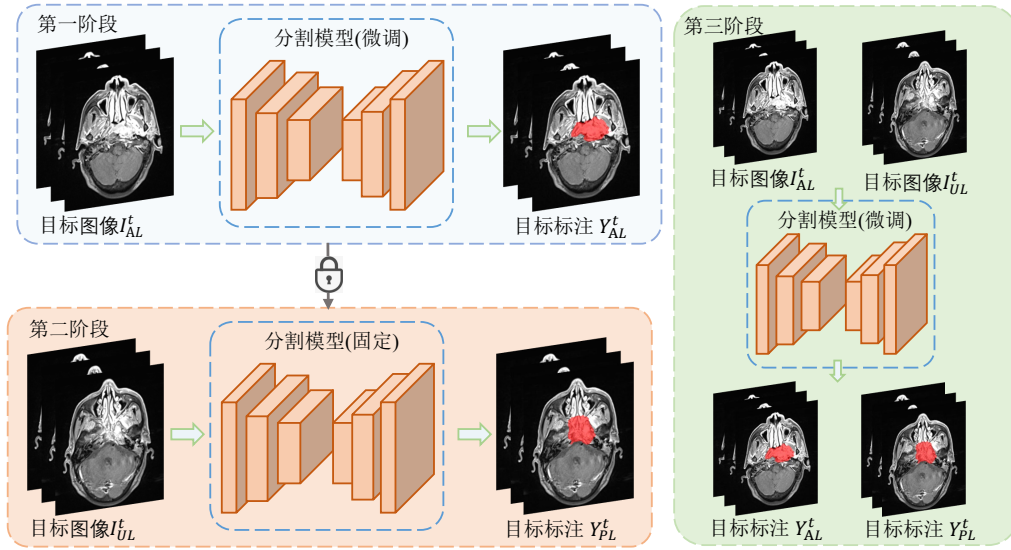


图 6-4 多阶段半监督学习算法流程

6.3 实验验证

6.3.1 实施细节

为了确保实验的公平性，本课题将每个中心数据随机划分为训练集，验证集和测试集，其比例为 7:1:2。其中每个中心的数据的 70% 被分配用于模型训练和参数调整。随后，选择在验证集上表现最佳的模型用于测试，并报告其在测试集上的结果。本实验首先在多中心鼻咽癌靶区勾画数据集上验证了提出方法在不同数据分布情况下的分割表现。考虑到较大数据集能提供更丰富的训练样本和更多样的语义信息，本文选择病人数量最多的 SMU 作为源域。随后，使用剩余四个中心的数据集作为无源域适应的目标域。此外，为了评估本文提出的方法的泛化能力和鲁棒性，进一步使用 SPH 中心的数据作为源域进行额外的验证性实验。随后，使用该方法将训练好的模型迁移到其他数据集上，以评估其性能和有效性。之后，本课题进一步在多标注者鼻咽癌靶区勾画数据集上验证该方法在不同标注风格差异性的鲁棒性。如多中心实验设置类似，本实验直接使用在 SPH 中心上训练好的模型来进行标注风格泛化。

考虑到不同医院的 MR 图像厚度差异巨大，本课题将该任务拆分为轴向上的二维图像分割，然后将分割结果重新恢复成三维预测结果。本实验使用 2D U-Net^[8] 作为所有方法的基础模型，并采用了基于 PyMIC^[217] 的开源实现。所有实验的训练和推理过程均在载有 2 张 NVIDIA RTX 3090 GPUs 的 Ubuntu20.04 服务器上进行。为了标准化训练过程，所有方法的模型输入都调整为统一分辨率的 256×256 ，并将强度归一化为零均值和单位方差。此外，为了提高模型的鲁棒性，在训练过

算法 6-1 无源主动域适应方法的训练过程

Data: 来自源域的图像及其标注 (I^s, Y^s) 和目标域的图像 I^t , 初始分割模型参数 θ

Result: 训练得到的最终的目标域图像分割模型

```

1 开始训练;
2 while 达到标注预算 do
3     根据公式6-1和来自源域的图像及其标注 ( $I^s, Y^s$ ) 训练得到源域模型  $\theta_s$ 
4     分别根据公式6-2, 公式6-3 和公式6-4获得源域数据的特征分布  $R^s$ 
5     根据公式6-2计算目标图像的特征  $F^t(I^t(i))$ 
6     根据公式6-5和公式6-6从目标域图像中选出  $\frac{p}{2}\%$  目标域代表性样本  $P^t$  和  $\frac{p}{2}\%$  源域代表性样本  $P^s$  组成  $I'_{AL}$ 
7     第一阶段, 根据公式6-7利用  $I'_{AL}$  进行模型微调, 得到微调后的目标域第一阶段模型  $\theta_t$ 
8     第二阶段, 利用目标域第一阶段模型  $\theta_t$  为所有未标注图像生成伪标签  $Y'_{PL}$ 
9     第三阶段, 根据公式6-7和主动标注样本及标注  $I'_{AL}$ ,  $Y'_{AL}$ , 未标注数据及其伪标签  $I'_{UL}$ ,  $Y'_{AL}$  来进行进一步的模型微调得到最终的目标域图像分割模型  $\theta_t$ 
10 end

```

表 6-1 U-Net 在不同数据集上鼻咽癌靶区分割性能的定量比较结果

数据中心	Dice (%)	95HD (mm)	ASD (mm)
SCH	72.19 ± 5.97	9.25 ± 3.52	2.22 ± 0.86
APH	86.79 ± 5.14	4.49 ± 3.35	1.13 ± 1.03
SPH	80.16 ± 8.09	4.42 ± 1.63	1.42 ± 0.71
WCH	85.87 ± 4.85	4.15 ± 2.47	1.17 ± 0.59
SMU	85.31 ± 7.45	5.88 ± 14.83	0.76 ± 0.69

程中对数据进行了随机旋转、翻转和加入高斯噪声进行数据增强。所有模型均使用 SGD 优化器和批量大小为 32 进行训练。根据消融研究, 聚类数 (表示为 K) 设置为 10。当使用源域或目标域中的所有数据进行训练时, 所有模型均进行 30000 次迭代训练。当使用主动选择的样本进行训练时, 数据规模较小, 因此进行 20000 次迭代训练。初始学习率设置为 0.03, 并且在每次迭代后以 0.9 的因子指数衰减。为了公平比较, 本实验使用相同的骨干 (U-Net) 重新实现所有比较方法, 并在相同的软硬件设置下比较。对于定量评估, 本实验评估了金标准和预测之间的 Dice 系数、95% 豪斯多夫距离 (95HD (mm)) 和平均表面距离 ASD (mm)。一般来说, 好的分割结果往往有更高的 Dice 分数和更低的 95HD 和 ASD 值。

表 6-2 SPH 和 SMU 到其他中心的直接预测结果的定量比较

域适应设置	Dice (%)	95HD (mm)	ASD (mm)
SPH $\rightarrow$ SCH	59.11 $\pm$ 8.98	31.35 $\pm$ 11.99	3.13 $\pm$ 1.73
SPH $\rightarrow$ APH	78.25 $\pm$ 9.42	9.14 $\pm$ 8.49	2.68 $\pm$ 2.95
SPH $\rightarrow$ WCH	74.76 $\pm$ 11.19	13.09 $\pm$ 12.78	4.02 $\pm$ 4.44
SPH $\rightarrow$ SMU	73.71 $\pm$ 13.55	12.89 $\pm$ 20.91	3.41 $\pm$ 5.92
SMU $\rightarrow$ SCH	63.35 $\pm$ 9.98	42.88 $\pm$ 17.29	2.49 $\pm$ 2.69
SMU $\rightarrow$ APH	81.45 $\pm$ 8.12	8.31 $\pm$ 8.97	2.60 $\pm$ 3.37
SMU $\rightarrow$ SPH	77.51 $\pm$ 8.49	9.42 $\pm$ 17.59	2.96 $\pm$ 4.95
SMU $\rightarrow$ WCH	77.40 $\pm$ 8.47	7.94 $\pm$ 7.07	2.37 $\pm$ 1.70

6.3.2 多中心鼻咽癌靶区无源主动域适应分割

1) 基线模型在多中心鼻咽癌靶区勾画数据集上的表现: 首先, 本实验调查了基线模型 2D UNet 在五个中心上各自单独训练和测试的表现。这些评估结果如表 6-1所示。实验结果表明, 即使是相同的模型和训练策略下, 不同的中心之间的分割结果存在很大的差异。具体来说, 在 SCH 的分割结果在 Dice 方面仅为 72.19%, 这可能是因为患者情况的更复杂和 SCH 中心的数据较少。相比之下, 其他四个中心的 Dice 均超过了 80%。此外, 实验结果还发现表明, 更大的数据集并不一定会产生更好的分割结果。例如, 在只有 146 名患者的 APH 中, 模型的分割表现优于具有 208 名患者的 SPH。

此外, 表 6-2还分别展示了将 SPH 和 SMU 数据上训练的分割模型, 直接用于其他中心测试时的分割性能。实验结果表明, 不同中心的数据差异导致模型在大多数数据集中 Dice 下降 5-10%, 例如 SPH $\rightarrow$ APH 和 SMU $\rightarrow$ SCH。此外还可以观察到, 有的数据集间, Dice 下降程度超过 10%, 例如 SPH $\rightarrow$ SCH 和 SPH $\rightarrow$ SMU。这进一步说明了在不同中心部署计算机辅助鼻咽癌靶区分割算法需要对模型进行主动的无源域适应。此外, 实验结果还发现从 SMU 训练的模型比从 SPH 训练的模型表现更好。这个实验结果可以给出一些启发, 如在实际应用中, 可以使用无源主动域适应方法将在具有更大数据集上训练的模型适应到小数据中心上, 以获得更好的分割结果。

2) 算法性能表现: 为了确保本章方法与现有相关工作的全面比较, 本实验不仅与域适应方法进行了比较(表6-3和6-4)还与最先进的主动学习方法进行了性能比较(表6-5和6-6)。在这两种比较下, 本实验报告了包含两种实验设置的广泛的实验结果(SPH $\rightarrow$ 其他数据集, SMU $\rightarrow$ 其他数据集)。

首先本章将提出的方法和最新的域适应方法进行了比较分析, 以全面评估

表 6-3 使用 SPH 数据集训练并适应到其他数据集上的定量结果。* 表示在所有配对的 t 检验中 p 值 < 0.05 ，表明该方法显著优于其他方法

	目标域	源域数据	Dice (%)	95HD (mm)	ASD (mm)
HistMatch ^[218]		No	57.84 $\pm$ 13.54	27.55 $\pm$ 12.15	4.33 $\pm$ 5.84
AdaptSeg ^[219]		Yes	59.63 $\pm$ 11.03	27.84 $\pm$ 17.52	7.31 $\pm$ 5.12
AdvEnt ^[204]		Yes	60.14 $\pm$ 10.16	20.49 $\pm$ 16.21	4.65 $\pm$ 2.72
UncertainDA ^[220]		No	61.74 $\pm$ 11.24	21.29 $\pm$ 16.91	4.84 $\pm$ 2.72
Tent ^[221]		No	62.37 $\pm$ 9.74	26.02 $\pm$ 17.27	5.45 $\pm$ 5.37
DPL ^[222]	SCH	No	62.88 $\pm$ 10.36	21.29 $\pm$ 14.56	4.27 $\pm$ 3.16
FSM ^[223]		No	62.85 $\pm$ 10.80	21.71 $\pm$ 18.17	5.42 $\pm$ 4.64
AdaMI ^[224]		No	62.24 $\pm$ 11.43	21.85 $\pm$ 19.63	5.32 $\pm$ 4.21
ProSFDA ^[225]		No	63.11 $\pm$ 9.75	21.68 $\pm$ 20.06	6.13 $\pm$ 5.99
CBMT ^[226]		No	62.74 $\pm$ 10.33	21.57 $\pm$ 18.06	6.09 $\pm$ 6.59
CPR ^[227]		No	63.42 $\pm$ 9.02	21.72 $\pm$ 19.69	5.81 $\pm$ 5.54
Ours		No	70.99 $\pm$ 6.59*	8.57 $\pm$ 2.39*	2.03 $\pm$ 0.84*
HistMatch ^[218]		No	77.82 $\pm$ 13.57	8.03 $\pm$ 8.76	2.54 $\pm$ 2.13
AdaptSeg ^[219]		Yes	77.87 $\pm$ 15.26	7.62 $\pm$ 7.19	1.97 $\pm$ 1.43
AdvEnt ^[204]		Yes	79.71 $\pm$ 11.37	7.13 $\pm$ 5.67	2.18 $\pm$ 1.63
UncertainDA ^[220]		No	80.03 $\pm$ 11.56	6.81 $\pm$ 5.22	2.01 $\pm$ 1.76
Tent ^[221]		No	62.37 $\pm$ 9.74	26.02 $\pm$ 17.27	5.45 $\pm$ 5.37
DPL ^[222]	APH	No	80.96 $\pm$ 12.14	6.23 $\pm$ 5.15	1.78 $\pm$ 1.49
FSM ^[223]		No	80.07 $\pm$ 8.62	6.25 $\pm$ 3.93	1.83 $\pm$ 1.28
AdaMI ^[224]		No	79.88 $\pm$ 11.81	6.09 $\pm$ 4.83	1.74 $\pm$ 1.39
ProSFDA ^[225]		No	80.97 $\pm$ 11.67	6.37 $\pm$ 3.95	1.72 $\pm$ 1.34
CBMT ^[226]		No	80.17 $\pm$ 12.31	5.90 $\pm$ 3.98	1.77 $\pm$ 1.51
CPR ^[227]		No	81.33 $\pm$ 10.60	6.05 $\pm$ 4.12	1.73 $\pm$ 1.14
Ours		No	85.69 $\pm$ 4.54*	4.58 $\pm$ 2.24*	1.18 $\pm$ 0.78*
HistMatch ^[218]		No	72.69 $\pm$ 10.84	10.50 $\pm$ 10.52	3.29 $\pm$ 2.67
AdaptSeg ^[219]		Yes	74.95 $\pm$ 8.64	10.74 $\pm$ 7.71	3.64 $\pm$ 2.45
AdvEnt ^[204]		Yes	75.27 $\pm$ 8.22	9.94 $\pm$ 6.69	3.19 $\pm$ 2.08
UncertainDA ^[220]		No	75.88 $\pm$ 8.81	12.14 $\pm$ 12.50	3.65 $\pm$ 4.02
Tent ^[221]		No	75.81 $\pm$ 10.31	11.24 $\pm$ 10.51	3.51 $\pm$ 3.57
DPL ^[222]	WCH	No	77.26 $\pm$ 9.83	8.96 $\pm$ 5.73	2.81 $\pm$ 1.55
FSM ^[223]		No	75.56 $\pm$ 8.58	10.30 $\pm$ 6.66	3.01 $\pm$ 1.68
AdaMI ^[224]		No	77.07 $\pm$ 9.78	8.35 $\pm$ 5.98	2.74 $\pm$ 1.42
ProSFDA ^[225]		No	77.81 $\pm$ 8.58	8.24 $\pm$ 5.45	2.93 $\pm$ 1.81
CBMT ^[226]		No	77.28 $\pm$ 9.54	8.47 $\pm$ 5.60	2.85 $\pm$ 1.38
CPR ^[227]		No	78.02 $\pm$ 8.85	8.14 $\pm$ 5.32	2.73 $\pm$ 1.31
Ours		No	83.59 $\pm$ 6.76*	4.85 $\pm$ 2.79*	1.37 $\pm$ 0.92*
HistMatch ^[218]		No	72.44 $\pm$ 17.93	9.70 $\pm$ 16.98	2.10 $\pm$ 4.59
AdaptSeg ^[219]		Yes	74.38 $\pm$ 12.58	12.05 $\pm$ 21.64	3.11 $\pm$ 4.97
AdvEnt ^[204]		Yes	75.56 $\pm$ 11.47	10.41 $\pm$ 19.32	2.79 $\pm$ 4.68
UncertainDA ^[220]		No	76.02 $\pm$ 12.67	9.45 $\pm$ 18.96	2.06 $\pm$ 4.57
Tent ^[221]		No	75.88 $\pm$ 12.15	8.17 $\pm$ 17.52	1.94 $\pm$ 3.14
DPL ^[222]	SMU	No	77.24 $\pm$ 10.53	9.22 $\pm$ 18.77	1.85 $\pm$ 3.23
FSM ^[223]		No	76.23 $\pm$ 14.25	9.93 $\pm$ 17.44	2.14 $\pm$ 3.63
AdaMI ^[224]		No	77.13 $\pm$ 14.88	8.79 $\pm$ 15.38	2.28 $\pm$ 2.75
ProSFDA ^[225]		No	77.65 $\pm$ 9.72	8.42 $\pm$ 15.13	2.09 $\pm$ 2.71
CBMT ^[226]		No	77.36 $\pm$ 9.84	9.29 $\pm$ 20.16	2.05 $\pm$ 3.09
CPR ^[227]		No	77.99 $\pm$ 9.77	8.77 $\pm$ 15.16	1.82 $\pm$ 2.65
Ours		No	84.13 $\pm$ 7.65*	6.17 $\pm$ 15.33*	0.86 $\pm$ 0.95*

表 6-4 使用 SMU 数据集训练并适应到其他数据集上的定量结果。* 表示在所有配对的 t 检验中 p 值 < 0.05 ，表明本方法显著优于其他方法

	目标域	源域数据	Dice (%)	95HD (mm)	ASD (mm)
HistMatch ^[218]		No	62.66 $\pm$ 10.49	41.66 $\pm$ 16.74	3.14 $\pm$ 4.94
AdaptSeg ^[219]		Yes	63.56 $\pm$ 10.57	41.38 $\pm$ 15.13	2.65 $\pm$ 2.73
AdvEnt ^[204]		Yes	64.06 $\pm$ 7.57	36.18 $\pm$ 14.12	2.90 $\pm$ 1.76
UncertainDA ^[220]		No	63.59 $\pm$ 8.32	28.73 $\pm$ 12.28	4.38 $\pm$ 2.85
Tent ^[221]		No	64.43 $\pm$ 7.55	29.25 $\pm$ 16.34	3.39 $\pm$ 2.23
DPL ^[222]		No	64.51 $\pm$ 7.70	24.74 $\pm$ 7.27	3.83 $\pm$ 1.85
FSM ^[223]		No	63.82 $\pm$ 9.15	22.65 $\pm$ 13.07	5.56 $\pm$ 3.44
AdaMI ^[224]		No	63.76 $\pm$ 7.32	25.86 $\pm$ 17.61	5.31 $\pm$ 4.68
ProSFDA ^[225]		No	64.05 $\pm$ 9.05	21.58 $\pm$ 14.41	4.35 $\pm$ 3.39
CBMT ^[226]		No	64.65 $\pm$ 9.17	26.71 $\pm$ 19.82	5.01 $\pm$ 3.96
CPR ^[227]		No	64.36 $\pm$ 10.79	20.24 $\pm$ 13.84	3.49 $\pm$ 2.39
Ours		No	72.58 $\pm$ 6.72*	8.06 $\pm$ 2.06*	2.00 $\pm$ 0.89*
HistMatch ^[218]		No	80.25 $\pm$ 9.35	7.88 $\pm$ 8.59	2.36 $\pm$ 2.96
AdaptSeg ^[219]		Yes	82.11 $\pm$ 7.15	8.26 $\pm$ 8.67	2.80 $\pm$ 2.21
AdvEnt ^[204]		Yes	80.91 $\pm$ 6.31	6.91 $\pm$ 5.93	2.33 $\pm$ 1.97
UncertainDA ^[220]		No	81.17 $\pm$ 7.11	7.96 $\pm$ 4.05	2.24 $\pm$ 1.28
Tent ^[221]		No	80.49 $\pm$ 7.58	7.94 $\pm$ 7.14	2.48 $\pm$ 2.56
DPL ^[222]		No	81.43 $\pm$ 6.74	6.38 $\pm$ 5.68	1.97 $\pm$ 2.02
FSM ^[223]		No	82.08 $\pm$ 12.40	6.94 $\pm$ 5.85	2.61 $\pm$ 1.77
AdaMI ^[224]		No	81.67 $\pm$ 8.52	6.04 $\pm$ 4.24	2.25 $\pm$ 1.84
ProSFDA ^[225]		No	82.29 $\pm$ 8.05	6.36 $\pm$ 4.93	2.34 $\pm$ 1.94
CBMT ^[226]		No	81.93 $\pm$ 8.35	6.52 $\pm$ 4.99	2.41 $\pm$ 1.96
CPR ^[227]		No	82.45 $\pm$ 8.37	6.05 $\pm$ 4.03	2.38 $\pm$ 1.95
Ours		No	86.01 $\pm$ 5.10*	4.38 $\pm$ 2.51*	1.21 $\pm$ 1.04*
HistMatch ^[218]		No	75.72 $\pm$ 12.88	8.06 $\pm$ 12.01	1.96 $\pm$ 1.81
AdaptSeg ^[219]		Yes	76.87 $\pm$ 9.62	8.85 $\pm$ 11.99	2.57 $\pm$ 2.07
AdvEnt ^[204]		Yes	76.71 $\pm$ 8.99	6.97 $\pm$ 11.29	1.90 $\pm$ 1.75
UncertainDA ^[220]		No	77.88 $\pm$ 9.58	6.44 $\pm$ 11.56	1.89 $\pm$ 1.79
Tent ^[221]		No	78.46 $\pm$ 9.15	7.37 $\pm$ 12.02	2.18 $\pm$ 2.44
DPL ^[222]		No	78.83 $\pm$ 8.38	6.89 $\pm$ 12.67	1.94 $\pm$ 2.95
FSM ^[223]		No	78.16 $\pm$ 7.91	7.12 $\pm$ 12.79	2.56 $\pm$ 2.76
AdaMI ^[224]		No	78.04 $\pm$ 8.57	7.32 $\pm$ 13.45	2.30 $\pm$ 2.26
ProSFDA ^[225]		No	78.61 $\pm$ 9.68	6.36 $\pm$ 6.78	1.97 $\pm$ 0.93
CBMT ^[226]		No	78.63 $\pm$ 7.84	7.04 $\pm$ 12.33	2.01 $\pm$ 2.66
CPR ^[227]		No	78.91 $\pm$ 8.51	7.29 $\pm$ 4.57	2.07 $\pm$ 1.99
Ours		No	80.97 $\pm$ 7.24*	4.02 $\pm$ 1.62*	1.03 $\pm$ 0.58*
HistMatch ^[218]		No	74.01 $\pm$ 15.16	7.56 $\pm$ 5.29	1.78 $\pm$ 1.29
AdaptSeg ^[219]		Yes	77.95 $\pm$ 8.57	7.39 $\pm$ 4.86	2.00 $\pm$ 1.19
AdvEnt ^[204]		Yes	78.37 $\pm$ 9.03	7.15 $\pm$ 4.84	1.95 $\pm$ 1.48
UncertainDA ^[220]		No	78.16 $\pm$ 9.10	6.67 $\pm$ 4.44	1.87 $\pm$ 1.07
Tent ^[221]		No	78.13 $\pm$ 8.79	6.46 $\pm$ 4.09	1.65 $\pm$ 0.99
DPL ^[222]		No	78.92 $\pm$ 8.42	6.26 $\pm$ 3.79	1.72 $\pm$ 1.04
FSM ^[223]		No	78.45 $\pm$ 7.45	6.61 $\pm$ 3.95	1.94 $\pm$ 1.18
AdaMI ^[224]		No	78.43 $\pm$ 9.16	6.37 $\pm$ 3.86	1.93 $\pm$ 0.96
ProSFDA ^[225]		No	79.04 $\pm$ 7.05	6.46 $\pm$ 3.67	1.86 $\pm$ 1.07
CBMT ^[226]		No	79.05 $\pm$ 7.14	6.72 $\pm$ 4.02	2.02 $\pm$ 1.18
CPR ^[227]		No	79.25 $\pm$ 7.07	6.59 $\pm$ 3.85	1.89 $\pm$ 1.28
Ours		No	84.71 $\pm$ 6.04*	4.48 $\pm$ 2.23*	1.24 $\pm$ 0.77*

表 6-5 SPH 数据集到其他数据集样本选择方法的定量结果。* 表示在所有配对的 t 检验中 p 值 < 0.05 ，表明本方法显著优于其他方法

	目标域	Dice (%)	95HD (mm)	ASD (mm)	NSD (%)
Random		62.28 $\pm$ 9.72	30.48 $\pm$ 10.59	2.64 $\pm$ 1.61	49.05 $\pm$ 6.37
Adversarial [219]		63.21 $\pm$ 7.23	24.75 $\pm$ 8.15	3.42 $\pm$ 1.61	49.47 $\pm$ 6.26
Entropy [204]	SCH	65.02 $\pm$ 8.67	18.03 $\pm$ 7.34	3.34 $\pm$ 1.87	54.45 $\pm$ 11.13
AADA [228]		65.08 $\pm$ 9.82	14.45 $\pm$ 5.90	2.88 $\pm$ 1.67	54.25 $\pm$ 8.87
MHPL [229]		67.35 $\pm$ 8.21	14.99 $\pm$ 8.61	3.41 $\pm$ 2.02	59.54 $\pm$ 9.15
STDR		69.55 $\pm$ 6.82*	10.64 $\pm$ 3.36*	2.83 $\pm$ 1.83	61.78 $\pm$ 10.05*
Random		81.19 $\pm$ 8.69	6.55 $\pm$ 4.82	2.01 $\pm$ 1.5	68.97 $\pm$ 12.44
Adversarial [219]		82.21 $\pm$ 6.32	8.27 $\pm$ 7.43	2.35 $\pm$ 1.81	69.40 $\pm$ 11.19
Entropy [204]	APH	83.45 $\pm$ 5.59	7.59 $\pm$ 9.45	1.82 $\pm$ 2.12	73.51 $\pm$ 12.03
AADA [228]		84.14 $\pm$ 5.05	7.87 $\pm$ 7.66	1.91 $\pm$ 1.95	73.34 $\pm$ 11.94
MHPL [229]		84.04 $\pm$ 5.17	5.77 $\pm$ 5.78	1.75 $\pm$ 1.42	74.02 $\pm$ 11.78
STDR		85.05 $\pm$ 5.25*	4.98 $\pm$ 3.07*	1.36 $\pm$ 1.01*	75.21 $\pm$ 11.86*
Random		78.05 $\pm$ 7.40	6.53 $\pm$ 6.20	1.92 $\pm$ 1.94	60.90 $\pm$ 10.95
Adversarial [219]		79.00 $\pm$ 7.87	7.07 $\pm$ 5.20	2.21 $\pm$ 1.42	62.68 $\pm$ 10.80
Entropy [204]	WCH	80.97 $\pm$ 7.98	7.50 $\pm$ 5.98	2.22 $\pm$ 1.37	66.79 $\pm$ 10.12
AADA [228]		81.21 $\pm$ 7.23	6.52 $\pm$ 3.81	2.01 $\pm$ 1.05	67.43 $\pm$ 11.45
MHPL [229]		80.40 $\pm$ 8.12	6.35 $\pm$ 4.74	1.81 $\pm$ 1.11	67.75 $\pm$ 13.38
STDR		82.81 $\pm$ 6.51*	5.39 $\pm$ 2.77*	1.58 $\pm$ 0.88*	71.17 $\pm$ 10.47*
Random		78.75 $\pm$ 10.26	7.36 $\pm$ 14.77	1.37 $\pm$ 2.45	68.79 $\pm$ 13.01
Adversarial [219]		79.44 $\pm$ 9.39	7.17 $\pm$ 12.57	1.58 $\pm$ 1.51	69.36 $\pm$ 12.63
Entropy [204]	WCH	80.27 $\pm$ 10.34	7.05 $\pm$ 14.74	1.43 $\pm$ 2.02	70.83 $\pm$ 12.54
AADA [228]		80.78 $\pm$ 8.60	6.33 $\pm$ 12.53	1.12 $\pm$ 1.62	71.47 $\pm$ 12.80
MHPL [229]		81.61 $\pm$ 8.32	7.76 $\pm$ 16.41	1.33 $\pm$ 1.09	73.36 $\pm$ 12.29
STDR		83.51 $\pm$ 8.05*	5.81 $\pm$ 13.10	1.26 $\pm$ 1.56	76.35 $\pm$ 12.22*

它们的性能差异。这个评估涵盖了有源域数据访问 (AdaptSeg [219], AdvEnt [204]) 和无源域数据访问 (HistMatch [218], UncertainDA [220], Tent [221], DPL [222], FSM [223], AdaMI [224], ProSFDA [225], CBMT [226] 和 CPR [227]) 的方法。正如预期的那样，表6-3和表6-4结果表明提出的方法与其他域适应方法相比取得了显著改进。这一结果强调了在半监督框架中与主动样本相结合，在提升性能方面起到了重要作用。例如，该方法在 SPH $\rightarrow$ SCH 和 SMU $\rightarrow$ SCH 上分别比其他方法取得的最佳 Dice 优势高出了 8.11% 和 8.07% 点。此外，图6-5展示了不同方法分割结果的可视化比较，与其他方法相比，本章方法的分割结果在二维和三维层面都与金标准重合度更高。

同时，考虑到该无源主动域适应框架本质上融入了主动学习的概念，将其与其他现有主动学习方法进行比较是合理的 [230]。因此，本实验选择将本章方法与随机样本选择以及其他四种最新的主动学习方法进行比较。具体如下：(i) 随机选择：从目标域中以均匀概率随机选择样本；(ii) Adversarial [219]：利用对抗学习中训练好的判别器选择目标样本的预测概率最低的样本，这表明这些样本在源域中与目标域之间的差异最显著；(iii) Entropy [204]：使用 AdvEnt 方法计算目标域内每个样本的预测熵，并选择熵最高的样本进行手动标注；(iv) AADA [228]：AADA

表 6-6 SMU 数据集到其他数据集样本选择方法的定量结果。* 表示在所有配对的 t 检验中 p 值 < 0.05 ，表明本方法显著优于其他方法

	目标域	Dice (%)	95HD (mm)	ASD (mm)	NSD (%)
SCH	Random	66.67 $\pm$ 7.51	15.06 $\pm$ 7.28	3.35 $\pm$ 2.23	59.31 $\pm$ 11.61
	Adversarial [219]	67.67 $\pm$ 7.20	19.85 $\pm$ 9.63	3.12 $\pm$ 2.31	59.45 $\pm$ 10.47
	Entropy [204]	67.78 $\pm$ 6.88	17.68 $\pm$ 8.55	4.77 $\pm$ 2.03	57.72 $\pm$ 10.07
	AADA [228]	67.96 $\pm$ 5.15	13.86 $\pm$ 9.73	2.53 $\pm$ 1.67	59.20 $\pm$ 8.29
	MHPL [229]	68.71 $\pm$ 7.59	12.15 $\pm$ 5.66	2.61 $\pm$ 1.85	60.26 $\pm$ 9.94
	STDR	70.99 $\pm$ 6.46*	10.72 $\pm$ 3.37*	2.47 $\pm$ 1.07	63.10 $\pm$ 8.84*
APH	Random	82.66 $\pm$ 6.37	7.28 $\pm$ 9.62	2.26 $\pm$ 1.79	71.74 $\pm$ 12.13
	Adversarial [219]	82.47 $\pm$ 7.23	7.16 $\pm$ 6.96	2.03 $\pm$ 1.94	72.03 $\pm$ 11.98
	Entropy [204]	83.24 $\pm$ 7.01	7.41 $\pm$ 8.46	2.15 $\pm$ 2.80	72.81 $\pm$ 12.61
	AADA [228]	83.21 $\pm$ 5.83	6.31 $\pm$ 7.03	1.83 $\pm$ 2.18	72.16 $\pm$ 11.71
	MHPL [229]	83.49 $\pm$ 8.94	5.65 $\pm$ 5.09	1.70 $\pm$ 1.79	73.79 $\pm$ 15.66
	STDR	85.49 $\pm$ 6.09*	4.63 $\pm$ 3.06*	1.24 $\pm$ 1.02*	76.01 $\pm$ 12.98*
SPH	Random	78.89 $\pm$ 8.79	6.66 $\pm$ 11.57	1.81 $\pm$ 1.88	64.74 $\pm$ 13.49
	Adversarial [219]	79.07 $\pm$ 7.79	9.08 $\pm$ 12.71	2.42 $\pm$ 2.51	66.16 $\pm$ 11.53
	Entropy [204]	79.11 $\pm$ 8.82	6.99 $\pm$ 11.75	1.85 $\pm$ 1.91	65.60 $\pm$ 13.67
	AADA [228]	79.77 $\pm$ 7.78	6.50 $\pm$ 5.79	1.77 $\pm$ 1.36	65.95 $\pm$ 12.99
	MHPL [229]	78.55 $\pm$ 7.43	5.86 $\pm$ 4.55	1.73 $\pm$ 1.12	65.59 $\pm$ 11.97
	STDR	80.00 $\pm$ 8.10	4.49 $\pm$ 2.64*	1.64 $\pm$ 1.46	66.36 $\pm$ 13.15
WCH	Random	79.77 $\pm$ 8.01	7.58 $\pm$ 3.95	2.38 $\pm$ 1.15	67.59 $\pm$ 11.83
	Adversarial [219]	81.18 $\pm$ 8.29	8.83 $\pm$ 10.41	2.83 $\pm$ 3.56	68.35 $\pm$ 11.47
	Entropy [204]	81.88 $\pm$ 7.38	6.03 $\pm$ 6.68	1.99 $\pm$ 2.74	69.74 $\pm$ 11.37
	AADA [228]	81.73 $\pm$ 8.35	5.70 $\pm$ 3.63	1.74 $\pm$ 1.44	68.89 $\pm$ 12.06
	MHPL [229]	82.92 $\pm$ 6.57	5.69 $\pm$ 3.54	1.61 $\pm$ 0.93	71.26 $\pm$ 12.83
	STDR	84.11 $\pm$ 6.64*	4.95 $\pm$ 2.75*	1.35 $\pm$ 0.88*	73.63 $\pm$ 11.46*

在进行样本选择时同时考虑样本的不确定性和多样性；(v) MHPL [229]: MHPL 选择同一分布中心内预测差异最大的样本作为代表性样本。为了公平比较，在所有比较实验中都使用相同的实验框架。每个实验都只选择 20% 的主动样本进行标记和训练。同时，本实验进一步使用了一个新的误差距离为 1mm 的归一化表面 Dice (Normalized Surface Distance, NSD) 来评估分割结果。表6-5和表6-6中的实验结果表明，设计的 STDR 策略选择的样本可以获得最佳的分割性能。这证明了所提出的策略的优越性，以及在样本选择中考虑目标域无关和目标域相关表示的有效性。

3) 消融实验: 为了验证提出方法的各个组成部分的有效性，本章在以下四种不同的设置下进行了全面的消融实验：(i) STDR- α : 选择在公式6-5中具有最大 $Simi$ 的 20% 样本，专注于目标域无关表示；(ii) STDR- β : 选择在公式6-5中具有最小 $Simi$ 的 20% 样本，专注于目标域相关表示；(iii) STDR: 按照公式6-6同时考虑两者；(iv) 提出的完整方法 (STDR+Semi)，将 STDR 策略与多阶段半监督学习相结合。如表6-7和6-8所示，STDR- α 和 STDR- β 在其对应的实验设定中均取得了良好的性能，相比而言，STDR- β 表现更好，这表明目标域的代表性样本对提高模型在目标域的分割性能是必要的。然而，当将 STDR- β 与 STDR 进行比较时，实验结果一致表明 STDR 总体上表现优于这两种设定 (p 值 < 0.05)。这表明，为

表 6-7 SPH 数据集到其他数据集的消融实验结果

	目标域	Dice (%)	95HD (mm)	ASD (mm)
STDR- α	SCH	62.66 $\pm$ 11.33	16.48 $\pm$ 16.03	4.54 $\pm$ 4.02
STDR- β		69.28 $\pm$ 7.78	14.93 $\pm$ 9.12	3.43 $\pm$ 2.71
STDR		69.55 $\pm$ 6.82	10.64 $\pm$ 3.36	2.83 $\pm$ 1.83
Ours (STDR+Semi)		70.99 $\pm$ 6.59	8.57 $\pm$ 2.39	2.03 $\pm$ 0.84
STDR- α	APH	81.46 $\pm$ 8.97	5.64 $\pm$.344	1.57 $\pm$ 1.19
STDR- β		84.38 $\pm$ 5.41	6.52 $\pm$ 7.48	1.89 $\pm$ 1.88
STDR		85.05 $\pm$ 5.25	4.98 $\pm$ 3.07	1.36 $\pm$ 1.01
Ours (STDR+Semi)		85.69 $\pm$ 4.54	4.58 $\pm$ 2.24	1.18 $\pm$ 0.78
STDR- α	WCH	79.51 $\pm$ 9.11	5.63 $\pm$ 3.23	1.58 $\pm$ 0.93
STDR- β		81.79 $\pm$ 7.31	6.06 $\pm$ 3.89	1.77 $\pm$ 0.97
STDR		82.81 $\pm$ 6.51	5.39 $\pm$ 2.77	1.58 $\pm$ 0.88
Ours (STDR+Semi)		83.59 $\pm$ 6.76	4.85 $\pm$ 2.79	1.37 $\pm$ 0.92
STDR- α	SMU	77.03 $\pm$ 13.11	8.45 $\pm$ 16.58	1.23 $\pm$ 1.34
STDR- β		82.66 $\pm$ 9.14	6.69 $\pm$ 15.83	1.17 $\pm$ 1.41
STDR		83.51 $\pm$ 8.05	5.81 $\pm$ 13.10	1.26 $\pm$ 1.56
Ours (STDR+Semi)		84.13 $\pm$ 7.65	6.17 $\pm$ 15.33	0.86 $\pm$ 0.95

了进一步提高模型性能，仅依赖于目标域相关的代表性样本是不够的。将目标域无关样本纳入考虑也是至关重要的，这样可以更有效地微调模型参数、保留共享的语义信息并提高模型的泛化能力，进而得到更准确的分割结果。最后可以观察到，本方法（STDR+Semi）在几乎所有指标上的性能都显著的超过了 STDR (p 值 < 0.05)，这为将半监督学习集成到主动域适应并取得准确的结果提供了有力证据。

为了进一步评估 STDR 策略的鲁棒性，实验了随机选择两种域迁移设定，并使用不同比例的主动样本进行了比较实验，如表6-9所示。为了消除由未标记样本和半监督学习策略可能带来的差异，本实验中仅使用主动标记的样本进行模型微调。结果如表6-9所示，可以观察到随着样本比例的增加，模型的性能表现出持续改善的趋势。还观察到在样本量较少时，引入新的主动标记样本时性能显著提升。然而，随着样本数量的增加，这种性能提升逐渐趋于平缓。因此，该方法选择 20% 的样本来进行标注以实现标注成本和分割性能之间的平衡。此外，与完全标记所有数据相比，本方法的性能在 SPH $\rightarrow$ SCH 和 SPH $\rightarrow$ APH 上 Dice 指标上只分别降低了 4.76% 和 2.34%。

此外，本研究还在 SPH $\rightarrow$ APH 设置下研究了聚类中心数 (K) 对分割结果的影响。本实验尝试了用不同数量聚类中心来选择主动样本，并在表6-11中展示了具体实验结果。可以观察到单个聚类中心并未产生最佳结果，可能是由于数据分布

表 6-8 SMU 数据集到其他数据集的消融实验结果

	目标域	Dice (%)	95HD (mm)	ASD (mm)
STDR- α	SCH	65.64 ± 7.84	16.02 ± 9.75	2.98 ± 1.55
STDR- β		68.21 ± 7.16	10.58 ± 3.35	2.99 ± 1.67
STDR		70.99 ± 6.46	10.72 ± 3.37	2.47 ± 1.07
Ours (STDR+Semi)		72.58 ± 6.72	8.06 ± 2.06	2.00 ± 0.89
STDR- α	APH	81.57 ± 7.89	6.78 ± 4.65	1.94 ± 1.16
STDR- β		83.57 ± 6.57	6.17 ± 8.25	1.93 ± 2.63
STDR		85.49 ± 6.09	4.63 ± 3.06	1.24 ± 1.02
Ours (STDR+Semi)		86.01 ± 5.10	4.38 ± 2.51	1.21 ± 1.04
STDR- α	SPH	76.80 ± 8.51	5.29 ± 2.56	1.67 ± 0.99
STDR- β		79.86 ± 7.97	5.42 ± 4.92	1.48 ± 1.21
STDR		80.00 ± 8.10	4.49 ± 2.64	1.64 ± 1.46
Ours (STDR+Semi)		80.97 ± 7.24	4.02 ± 1.62	1.03 ± 0.58
STDR- α	WCH	79.44 ± 8.27	6.48 ± 4.97	1.68 ± 1.12
STDR- β		83.55 ± 6.69	6.53 ± 5.33	1.70 ± 1.14
STDR		84.11 ± 6.64	4.95 ± 2.75	1.35 ± 0.88
Ours (STDR+Semi)		84.71 ± 6.04	4.48 ± 2.23	1.24 ± 0.77

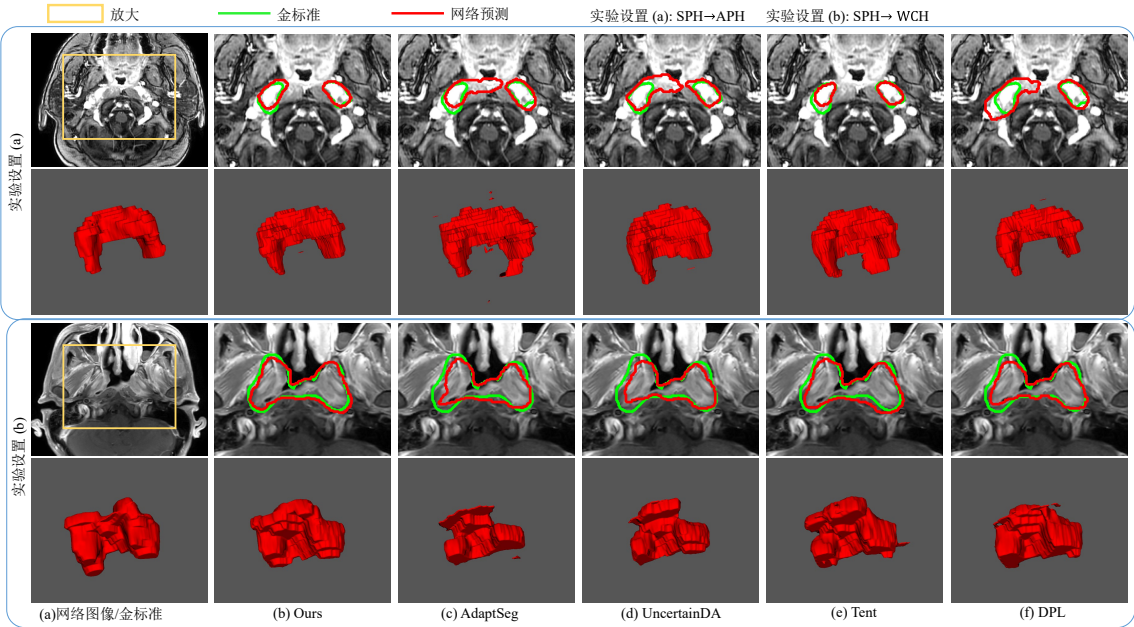


图 6-5 不同域适应方法的可视化比较。第一行和第三行展示了二维的可视化结果，第二行和第四行提供了三维结果

的复杂特性，这可能是由供应商、扫描参数等差异导致的。最终实验选择了使用 10 作为聚类中心数，因为这种设置训练的模型在测试中表现最为优异。

表 6-9 不同主动样本数量下，本章方法的多中心泛化结果

域适应设置	标注比例	Dice (%)	95HD (<i>mm</i>)	ASD (<i>mm</i>)
SPH → SCH	10%	64.61 ± 9.59	11.53 ± 4.21	2.96 ± 1.67
SPH → SCH	20%	69.55 ± 6.82	10.64 ± 3.36	2.83 ± 1.83
SPH → SCH	40%	70.88 ± 6.75	9.96 ± 4.49	2.47 ± 1.21
SPH → SCH	60%	71.73 ± 6.79	9.22 ± 4.28	2.66 ± 1.62
SPH → SCH	80%	73.57 ± 5.59	8.82 ± 4.29	2.27 ± 1.19
SPH → SCH	100%	74.31 ± 5.57	8.57 ± 4.68	2.09 ± 1.01
SPH → APH	10%	84.05 ± 5.60	5.37 ± 4.24	1.48 ± 1.35
SPH → APH	20%	85.05 ± 5.25	4.98 ± 3.07	1.36 ± 1.01
SPH → APH	40%	85.62 ± 4.76	4.98 ± 3.13	1.35 ± 1.07
SPH → APH	60%	86.02 ± 4.39	4.60 ± 2.47	1.33 ± 0.87
SPH → APH	80%	86.81 ± 4.19	4.05 ± 1.99	1.18 ± 0.93
SPH → APH	100%	87.39 ± 4.74	3.97 ± 2.19	0.99 ± 0.75

4) 多标注者鼻咽癌靶区无源域适应分割：上一节在多中心鼻咽癌靶区勾画数据集上全名的比较和分析了本章提出方法的有效性，本节旨在评估该方法在多标注者鼻咽癌靶区勾画数据集上的通用性。和上节实验设置类似，本实验以各个标注者自己标注结果训练和测试结果作为模型上限结果 (Upper bound) 以直接用源域模型测试结果作为下限结果 (Lower bound)，并将提出的方法与一系列最新的方法做对比。为了同时考虑方法的通用性和鲁棒性，本实验采用了另一个常有的有限标注数据的数据标注比例 10%，并研究了从 SPH 中心训练模型到不同标注风格泛化性能。表6-10展示了这些方法的定量结果比较，可以发现本章提出的方法也可以在标注风格存在差异的情况下取得较大的实验结果，并超越了已有的域适应和主动学习方法。图6-6展示不同域适应方法对同一病人的分割结果，可以观察到，本文提出的方法二维和三维图像上的预测与金标准重合度更好。这些结果再次证明了本章方法是一个有效且通用的方法。

6.3.3 临床可用性分析

此外，为了研究本章方法在真实临床场景下的可用性，本课题进一步开展了专家评估和临床相关性分析。遵循临床评估流程，首先在跨中心适应（从 SMU 到其他数据集）设置中，分别从 APH 和 SPH 的测试集中随机选择了 20 名患者的预测结果用于专家评估。然后，一名肿瘤科主任医生审查和修改了这 40 名患者的错

表 6-10 SPH 数据集训练的模型适应各个标注专家结果的定量分析。* 表示在所有配对的 t 检验中 p 值 < 0.05 ，表明该方法显著优于其他方法

	Dice (%)				95HD (mm)			
	专家 1	专家 2	专家 3	平均	专家 1	专家 2	专家 3	平均
Lower bound	70.79 $\pm$ 16.89	63.95 $\pm$ 14.28	65.19 $\pm$ 14.95	65.73 $\pm$ 15.06	15.69 $\pm$ 13.85	17.97 $\pm$ 11.48	18.25 $\pm$ 14.00	16.83 $\pm$ 12.10
Upper bound	83.15 $\pm$ 7.85	75.87 $\pm$ 7.95	74.83 $\pm$ 13.58	78.08 $\pm$ 9.03	5.92 $\pm$ 3.83	7.20 $\pm$ 4.39	7.39 $\pm$ 6.14	6.53 $\pm$ 4.17
AdaptSeg ^[219]	70.93 $\pm$ 13.86	64.57 $\pm$ 12.44	65.69 $\pm$ 14.66	66.14 $\pm$ 13.08	14.24 $\pm$ 13.20	15.73 $\pm$ 11.12	17.61 $\pm$ 13.83	15.45 $\pm$ 11.52
AdvEnt ^[204]	71.22 $\pm$ 16.55	65.12 $\pm$ 11.17	66.21 $\pm$ 11.79	66.57 $\pm$ 12.93	15.02 $\pm$ 13.72	14.88 $\pm$ 10.47	16.36 $\pm$ 12.31	15.11 $\pm$ 11.23
UncertainDA ^[220]	71.57 $\pm$ 15.11	65.52 $\pm$ 11.81	66.41 $\pm$ 13.38	66.89 $\pm$ 12.65	13.86 $\pm$ 11.57	15.13 $\pm$ 9.86	15.54 $\pm$ 11.38	14.59 $\pm$ 10.52
Tent ^[221]	72.13 $\pm$ 15.25	65.93 $\pm$ 11.88	66.75 $\pm$ 13.34	67.32 $\pm$ 12.88	13.58 $\pm$ 11.52	14.95 $\pm$ 9.77	15.34 $\pm$ 12.77	14.28 $\pm$ 10.40
DPL ^[222]	72.59 $\pm$ 14.15	66.13 $\pm$ 10.08	66.92 $\pm$ 13.28	67.55 $\pm$ 12.32	12.85 $\pm$ 12.58	13.77 $\pm$ 8.48	14.84 $\pm$ 11.41	13.60 $\pm$ 10.10
FSM ^[223]	71.88 $\pm$ 14.03	65.97 $\pm$ 11.47	66.60 $\pm$ 15.62	67.20 $\pm$ 12.98	13.92 $\pm$ 13.15	14.54 $\pm$ 10.1	14.43 $\pm$ 8.46	13.83 $\pm$ 9.60
AdaMI ^[224]	72.45 $\pm$ 13.12	66.39 $\pm$ 11.04	67.13 $\pm$ 12.64	67.73 $\pm$ 12.15	13.09 $\pm$ 11.48	13.56 $\pm$ 8.98	14.18 $\pm$ 11.75	13.33 $\pm$ 9.90
ProSFDA ^[225]	73.22 $\pm$ 13.55	66.98 $\pm$ 10.93	67.48 $\pm$ 12.07	68.29 $\pm$ 12.05	13.02 $\pm$ 10.72	12.67 $\pm$ 7.06	12.56 $\pm$ 9.89	12.48 $\pm$ 8.60
CBMT ^[226]	73.09 $\pm$ 14.27	67.26 $\pm$ 9.71	67.76 $\pm$ 11.65	68.53 $\pm$ 11.37	12.62 $\pm$ 12.99	12.17 $\pm$ 6.93	12.66 $\pm$ 10.36	12.46 $\pm$ 9.47
CPR ^[227]	73.42 $\pm$ 13.35	67.75 $\pm$ 10.77	67.93 $\pm$ 12.19	68.98 $\pm$ 11.32	12.26 $\pm$ 12.51	11.78 $\pm$ 6.69	12.31 $\pm$ 9.41	11.92 $\pm$ 8.56
Random	74.66 $\pm$ 10.61	67.97 $\pm$ 9.96	68.27 $\pm$ 13.17	70.14 $\pm$ 10.35	9.67 $\pm$ 8.78	10.39 $\pm$ 5.69	10.63 $\pm$ 8.43	10.17 $\pm$ 6.97
Adversarial ^[219]	74.30 $\pm$ 10.74	68.52 $\pm$ 10.24	69.86 $\pm$ 13.89	70.77 $\pm$ 11.03	7.99 $\pm$ 4.30	10.81 $\pm$ 6.01	10.57 $\pm$ 8.06	9.97 $\pm$ 6.16
Entropy ^[204]	76.62 $\pm$ 10.13	68.93 $\pm$ 9.98	69.25 $\pm$ 13.31	71.33 $\pm$ 10.73	9.45 $\pm$ 8.41	10.66 $\pm$ 6.40	9.93 $\pm$ 6.25	10.10 $\pm$ 7.04
AADA ^[228]	77.51 $\pm$ 10.08	70.33 $\pm$ 8.53	70.08 $\pm$ 12.29	72.33 $\pm$ 9.92	8.49 $\pm$ 6.61	11.09 $\pm$ 5.54	9.94 $\pm$ 8.05	9.51 $\pm$ 6.33
MHPL ^[229]	78.15 $\pm$ 8.61	71.05 $\pm$ 8.47	70.43 $\pm$ 12.30	73.01 $\pm$ 9.56	7.87 $\pm$ 7.06	10.56 $\pm$ 5.91	9.17 $\pm$ 5.42	8.82 $\pm$ 5.59
Ours	79.76 $\pm$ 6.11*	72.13 $\pm$ 7.94*	71.70 $\pm$ 11.25*	74.25 $\pm$ 8.22*	7.77 $\pm$ 5.63	10.49 $\pm$ 6.64	8.21 $\pm$ 5.42*	8.47 $\pm$ 5.45

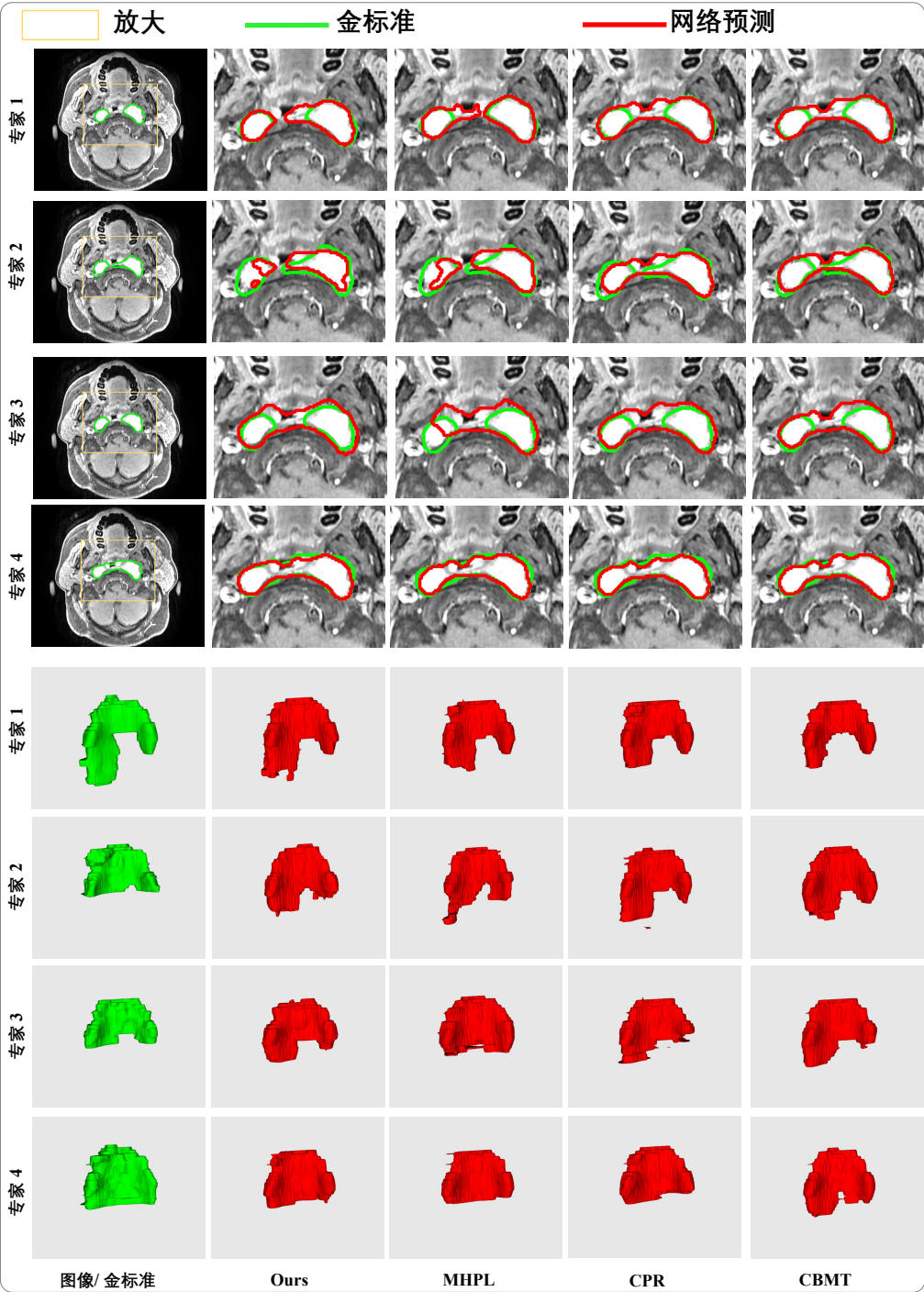


图 6-6 不同域适应方法在多标注者分割数据集上的可视化比较。最前四行展示了二维切片的预测轮廓，后四行展示了三维区域预测结果

表 6-11 SPH 到 APH 实验设置下, STDR 方法中聚类数量对实验结果的影响

聚类中心数 (K)	Dice (%)	95HD (mm)	ASD (mm)	NSD (%)
1	83.81 $\pm$ 5.93	7.31 $\pm$ 8.46	2.06 $\pm$ 2.51	72.91 $\pm$ 12.31
5	84.41 $\pm$ 5.37	5.80 $\pm$ 6.67	1.56 $\pm$ 1.84	73.84 $\pm$ 12.32
10	85.05 $\pm$ 5.25	4.98 $\pm$ 3.07	1.36 $\pm$ 1.01	74.73 $\pm$ 11.77
20	84.71 $\pm$ 5.04	5.97 $\pm$ 6.57	1.56 $\pm$ 1.71	74.42 $\pm$ 11.52
30	84.69 $\pm$ 5.01	5.43 $\pm$ 4.36	1.52 $\pm$ 1.26	74.62 $\pm$ 11.36

误预测直到结果临床可用。最后, 将模型的预测与其对应的专家修正结果进行比较, 通过 $100 \times (1 - \text{Dice})\%$ 来计算修改程度。本课题进一步调查了方法在处理不同标注风格时结果的临床可用性, 该过程随机从多标注者数据的测试集中选择了 20 名病人的预测结果, 并邀请当初分别这些数据的专家 (专家 1、2、3、4) 审查和修改预测结果直到结果临床可用。本实验也采用相同的方法计算每个预测的修改程度。专家评估结果如表 6-12 所示。结果表明, 在 SPH 和 APH 数据集上, 80% 的患者和 95% 的患者的修改量不超过 10%。SPH 和 APH 的平均修改量分别为 6.83% 和 5.52%。对于不同标注风格任务, 专家修改程度均不超过 10%。而且所有病人的平均修改时间不超过 110s, 远低于医生手工勾画靶区耗费的时间 (平均每个病人需要 10 到 15 分钟)。

同时, 本课题进一步研究了肿瘤分期 (T 分期) 对模型性能的影响, 因为不同 T 分期的患者具有不同的肿瘤大小和边缘及影像学表现, 这些差异与临床治疗息息相关。因此, 本文在具有不同 T 分期的患者分割结果中进行了定量比较。考虑到测试队列的样本量, 本研究选择了 SPH 和 WCH 数据集进行实验, 其结果如图 6-7 所示。在从 SMU 到 SPH 的实验设置中, T1+2 患者和 T3+4 患者的平均 Dice 分别为 78.53% 和 82.57%; T1+2 患者和 T3+4 患者的平均 95HD 分别为 4.12 mm 和 3.96 mm; 平均 ASD 分别为 0.98 mm 和 1.07 mm。两个亚组之间的 Dice, 95HD, ASD 均没有明显差异 (p -值 > 0.05)。在从 SMU 到 WCH 的实验设置中, T1+2 患者和 T3+4 患者的平均 Dice 分别为 84.92% 和 84.62%; T1+2 患者和 T3+4 患者的平均 95HD 分别为 4.04 mm 和 4.87 mm; 平均 ASD 分别为 1.09 mm 和 1.38 mm。同样, 在两个亚组之间观察到的 Dice, 95HD, ASD 均没有明显差异 (p -值 > 0.05)。这些结果说明本方法对于 T 分期不敏感。综上, 这些实验结果说明了该方法可以在有限的医生介入下达到临床可用的程度, 证明了该方法的有效性和实用性。

表 6-12 不同数据集中专家修正程度的定量结果

	修改程度 (%)				平均修改程度 (%)	修改耗时 (s)
	0-5%	> 5%-10%	> 10%-15%	> 15%		
SPH	7 (35%)	9 (45%)	4 (20%)	0 (0%)	6.83 ± 3.41	108.5 (74.5-131.5)
APH	12 (60%)	7 (35%)	1 (5%)	0 (0%)	5.52 ± 2.85	93.0 (75.2-106.8)
专家 1	8 (40%)	7 (35%)	3 (15%)	2 (10%)	7.3 ± 4.76	108.0 (86.0-168.0)
专家 2	7 (35%)	9 (45%)	1 (5%)	3 (15%)	8.23 ± 6.18	100.0 (87.3-114.8)
专家 3	10 (50%)	7 (35%)	2 (10%)	1 (5%)	6.57 ± 7.05	75.5 (64.5-116.6)
专家 4	12 (60%)	7 (35%)	1 (5%)	0 (0%)	3.75 ± 3.33	88.0 (40.8-153.8)

6.4 讨论与分析

精确的肿瘤分割对于癌症患者的精准放疗至关重要^[231,232]。尽管医学图像分割算法的数量不断增加^[233-238]，但它们的实际应用面临着重大挑战，尤其是在新的医疗中心应用时，源域训练的模型会出现显著的性能下降。结合表6-1和表6-2，可以清楚地看到这些算法在 Dice 方面一致表现出 5% 至 10% 甚至超过 10% 的降低，在 95HD 和 ASD 指标上也有明显的性能下降。例如，在适应 SMU $\rightarrow$ SCH 或 SPH $\rightarrow$ SCH 的情况下，95HD 显示的误差变化超过 20 mm，表明了显著的性能差异。此外，医疗数据的隐私和安全性以及医院规定也限制了多中心模型的开发。即使某些医疗中心拥有大量标记良好的数据，其他医疗中心也无法利用这一资源，主要是由于不同数据域之间明显的领域差距和数据隐私造成的。这些数据之间的分布差异在图2-14中通过特征可视化和灰度分布曲线的形式展现了出来。

这些挑战为新的域适应技术开发以弥合域差异指明了方向，为此本文提出了一种新的解决方案，即无源主动领域适应。该方法相对于现有的域适应方法具有多个优势。首先，它只需要有限数量的参考向量，无需源数据访问。这不仅确保了数据的隐私和安全性，还便于信息传输和重复利用。其次，本方法提升了模型性能。如表6-3和表6-4所示，以往的方法确实提高了目标领域的分割模型性能，但它们完全依赖于未标记的目标数据。缺乏监督训练的金标准可能会引入模型过拟合，导致相对较小的性能提升，通常在 Dice 的 1-3% 范围内。此外，在实际临床场景中，对有限数据进行主动标注在时间开销和标注成本上既可行又相对高效。本方法的主要意义在于其利用有限标注数据的能力，使域适应模型能够达到与特定于目标域的全监督训练模型相当的性能水平。这凸显了该方法潜在的实际应用价值。例如，结合表6-1、6-3和6-4，可以发现在 SPH $\rightarrow$ SCH 中，该方法的性能与直接在 SCH 上训练的结果可比，在 Dice 方面取得了差不多的结果，在 95HD 和 ASD 方面取得了更好的结果。同时，在 SMU $\rightarrow$ SCH 中，该方法明显优于其他方法。另

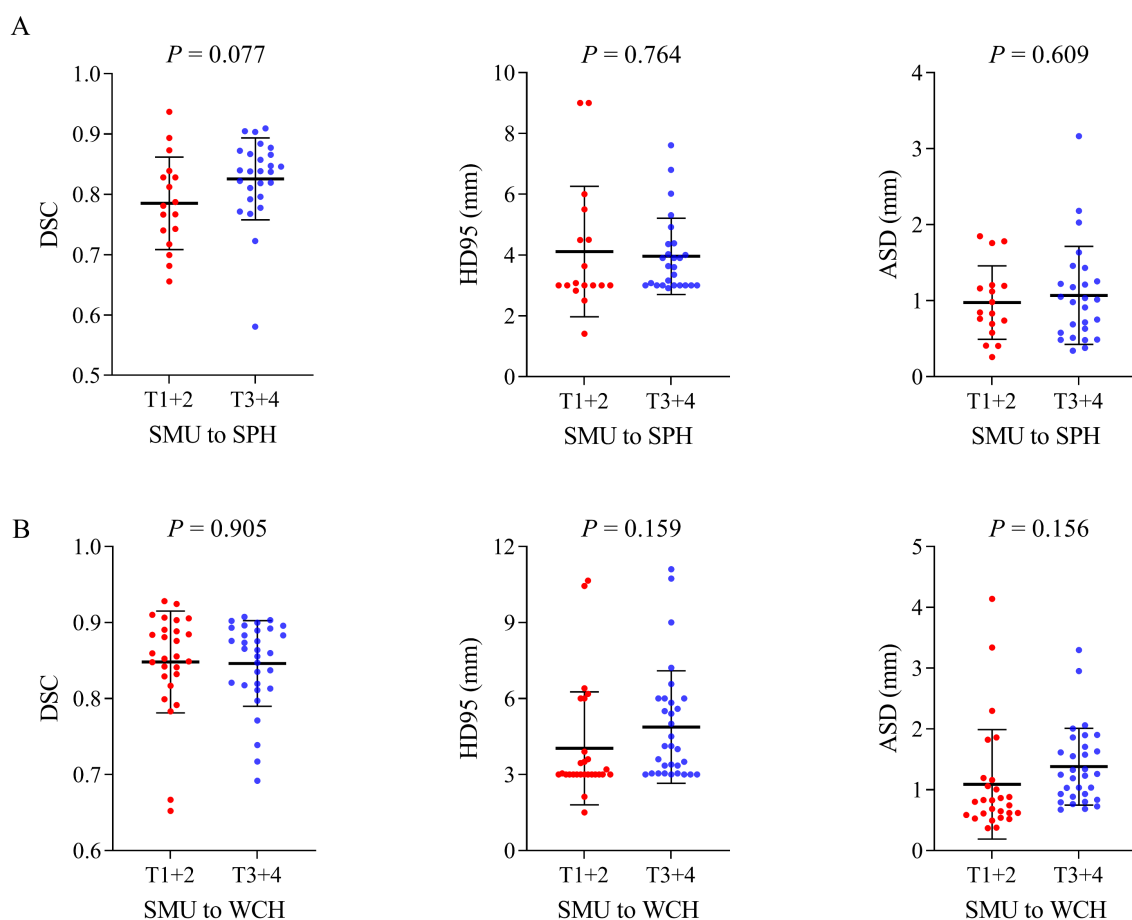


图 6-7 SMU 到 SPH 和 WCH 时，肿瘤不同 T 分期之间的准确性比较

外，专家评估结果也说明了本方法可以在较低的标注成本下实现临床可以的分割效果，能大幅度减少医生勾画时间，加速放疗流程。

尽管如此，这项工作在临床应用和技术创新方面也存在一些局限性。首先，由于 MRI 具有较好的软组织对比度，本文模型的开发和验证都是使用 MRI 图像进行的。此外，CT 更常用于治疗计划制定，但通常的做法是在 MRI 上进行靶区勾画，然后通过图像配准将其配准到 CT 上。因此，如果能同时利用配对的 MRI 和 CT 数据集进行实验验证，能更好的促进临床应用。另外，使用欧氏距离作为 K-Means 聚类中的度量指标在低维特征中表现良好，但在高维表示中可能会导致偏差^[239]。虽然本方法的特征表示维度是 1×256 ，与由数千个 MR 样本够成的整个数据集相比（除 SCH 外）相比并不高^[240]，但在这方面仍可能存在优化的空间，可以采用其他度量标准。此外，多阶段半监督学习依赖于第一阶段训练后生成的伪标签的准确性。这些伪标签可能存在错误，但如表6-7和表6-8所示，本方法仍然显示出了性能提升。然而，本文目前并未探讨处理不正确标签的策略，这可能会限制模型性能。此外，这项工作仅在多中心和多标注者鼻咽癌靶区勾画 MR 数据集上进行了

评估，缺乏对其他任务和数据模态的鲁棒性和泛化性评估。将此方法应用于其他不同模态和不同疾病的临床任务也有利于进一步推动该话题的进展。未来，本课题将继续围绕基于有限标注的无源主动域适应探索更高效和鲁棒的深度学习方法。

6.5 本章总结

在本章中，介绍了一种新的无源主动领域适应方法，用于跨中心医学图像分割，并在两个多中心和多标注者鼻咽癌靶区勾画数据集上的 1227 名患者数据上进行了严格验证。首先，本章设计了一种 STDR 策略，以区分目标不同中心数据中的目标域无关和目标域相关代表性样本。这个过程通过样本特征聚类来筛选代表性样本，并且使用这些标记的代表性样本进行模型训练进而提高模型在目标域上的性能。此外，本章还开发了一个多阶段半监督学习方法，将其与主动域适应相结合，充分利用目标域中剩余的未标注样本来增强模型分割效果。大量的实验结果表明，本章方法于最新的域适应和主动学习方法。与此同时，本章方法在较小的标注成本情况下达到了接近全监督模型的性能。最后，专家评估表明，本章的方法可以在很少的用户介入下达到临床可用的效果，能有效减轻医生的标注负担，临床应用前景明朗。在未来，本课题计划考虑在其他分割任务上验证该方法（例如不同的成像模态和不同病种）。此外，本课题还将尝试用不确定性修正策略来校正伪标签进而优化多阶段半监督学习，以提升模型的分割性能。

第七章 总结与展望

7.1 全文总结

针对医学图像分割任务中深度学习模型依赖大规模高质量标注数据，但医学图像标注依赖临床经验、耗时长、成本高等问题；本文提出了一系列基于有限标注数据的医学图像分割方法降低标注成本的同时提高模型的分割性能。针对有限标注学习中数据标注四种主要形式（人工交互、稀疏标注、少量标注和主动标注），本文从四个角度设计了四种结合深度学习的有限标注方法，包括基于深度学习的交互式分割方法、基于稀疏标注的弱监督学习、结合少量有标注数据和大量无标注数据的半监督学习、以及选取最有信息量样本进行标注的主动学习。在多个公开基准数据集和本文收集的多中心临床数据集上大量实验验证了本文提出方法在分割性能、数据利用效率上的优异表现，为开发低标注成本、高性能的医学图像分割模型提供了可行的解决方案。此外，临床评估也表明了本文提出方法只需要极少的医生介入即可实现临床可用，可以大幅加速临床分割流程及降低医生手工勾画负担。

7.2 论文工作创新点

为了减少高性能模型对标注数据的依赖，本论文从四个角度出发针对不同类型的有效标注数据和方法，提出了一系列的模型训练方法来降低数据标注成本的同时提高模型的性能表现，并进行实际临床验证促进方法落地。本论文的主要创新点如下：

1. 针对已有基于深度学习的交互式分割方法需要大量用户交互且难以分割未见过图像的不足，提出了基于测地线距离变换的可泛化交互式分割标注，从用户交互形式、交互信息编码形式和交互式结果修正流程三个角度来提高交互式分割方法的可用性、准确性和通用性。多中心多模型数据集上的实验结果表明，提出的交互式分割方法可以在有限的用户交互下实现准确的图像分割，并可以进一步用于标注训练时没见过的图像，证明了 MIDeepSeg 的有效性和通用性。

2. 针对深度学习模型训练过程中涂鸦标注提供的监督信号有限的问题，本文通过结合双分支网络（Dual-Branch Network, DBN）和伪标签方法，提出了一种动态混合伪标签的弱监督学习训练策略（Dynamically Mixed Pseudo Labels Supervision, DMPLS），能充分利用涂鸦标注来生成高质量伪标签提升分割网络的性能。通过在心脏结构分割 MRI 数据集和腹部多器官 CT 数据集上的大量实验表

明，该方法可以利用有限的涂鸦标注提升模型的分割性能，并显著的优于现有最先进的弱监督方法。

3. 针对临床实践中收集有标注数据困难但收集无标注数据简单的现状，本文通过融合水平集方法和深度学习，提出了一种双任务一致性半监督学习算法 (Dual-Task Consistency, DTC)，以充分利用大量无标注数据来训练高性能分割模型。在左心房分割 MRI 数据集和腹部胰腺分割 CT 数据集上的大量实验表明，本文方法可以有效利用无标注数据提升模型性能，并优于现有的半监督医学图像分割算法。

4. 针对现有有限标注学习方法需要数据属于同一分布且未考虑图像间的信息量差异等问题，本文通过融合主动学习和域适应方法提出了一个新的有限标注学习场景，无源主动域适应 (Source-Free Active Domain Adaptation, SFADA)，通过引入目标域上的有限标注数据来提高源域训练模型在目标域上的分割表现。为了降低数据标注成本和提升模型性能，主动学习被用于筛选信息量最大的样本进行标注，并利用半监督学习进行模型微调。本文首次从五家医院收集 1227 个鼻咽癌病人的核磁数据并构建了两个大规模的多中心、多标注者鼻咽癌靶区勾画数据集。在这两个临床数据集上的大量实验结果表明，本文提出的方法利用有限标注数据可以很好地适应不同数据分布和不同勾画风格的分割任务。此外，多中心临床验证进一步证明了该方法只需要少量的医生介入即可达到临床可用水平。

综上所述，通过方法分析与实验验证的紧密结合，本论文为利用有限标注数据进行模型训练这一问题，提供了一系列创新性解决方案，加速了算法模型在临床的应用过程。这些研究成果不仅为技术研究提供了新的视角和思路，也为临床实践产品的开发提供了解决方案，为基于深度学习的医学图像分割模型在临床场景下的大规模应用奠定了坚实基础。

7.3 未来工作展望

尽管本课题所提出的方法在一定程度上降低了医学图像分割算法开发的标注成本，但仍存在一些问题和挑战需要进一步研究和解决。简要总结如下：

(1) 在目前的交互式分割工作中，既往工作以及本文提出的 MIDeepSeg 都主要聚焦于分割单个器官或病灶。但在实际临床中，多类分割才是主要是主要分割任务，如图2-11所示的腹部多器官分割任务。此外，这些方法主要集中于解决单个结构组成的器官或病灶，如肝脏、颅脑胶质瘤等，对于多个小病灶分割任务来说，如图7-1所示，如何设计有效的交互流程是今后研究需要关注的重点。

(2) 在当前的弱监督医学图像分割任务中，涂鸦标注因其能更好地表示分割

目标的解剖信息成为了最主要的稀疏标注形式，但相比于点、框和分类信息来说，涂鸦的获取成本相对更高，具体差异如图1-4所示。未来如何利用成本更低的点、框或分类信息来达到和涂鸦标注相比差不多或更好的分割性能也是研究的重点。

(3) 第五章提出的半监督学习方法只在单类分割任务上进行了验证，如何用水平集或其他分割表示方法来表示多类分割任务依然值得探索。此外，如何将该方法扩展到其他像素级预测任务也是值得关注的问题，如边缘提取^[210]和关键点估计^[211]，即如何设计跨任务的转换和一致性约束。未来，希望此外该方法可以引起计算机视觉社区对不同任务一致性的关注，因为在许多方向上都可以以一种半监督的方式一致地构造任务，例如两流视频识别^[212]、多任务图像重建^[213,214]、图像分割和边缘检测等，以利用大量未标记数据。

(4) 第六章提出的无源主动域适应方法验证了利用主动学习来选择有限样本进行标注对缓解深度学习跨域时模型性能降低问题有较大帮助。因此，后续研究可以更深入地探讨如何针对该任务设计更有效的样本选取策略，如扩展2.2.4提及的采样方法。如何针对这些信息量高的样本进行标注，采用什么样的标注形式成本更低，如2.2.2讨论过的标注策略，也是有意义的研究问题。此外如何利用这些有限的标注数据进行模型微调也是重要的研究方向，例如探索使用2.2.3中提及的方法等。

(5) 目前本研究所有方法均为基于卷积神经网络的深度学习方法，Transformer^[67]已经逐步成为了新一代深度学习方法的基础网络架构。因此，未来可以考虑使用Transformer或融合CNN与Transformer来推进有限标注学习的发展，进一步开发出高性能低标注成本的深度学习模型。

(6) 为了降低标注成本，本文主要是从利用有限的标注信息来开发模型，目前自监督学习^[68,241]利用图像恢复、图像对比、图像补全和代理任务预测等训练更好的初始化模型也获得了广泛的关注。因此，未来也可以考虑将现有的有限标注学习方法和自监督学习结合起来，利用自监督学习获得的预训练模型在下游任务的有限标注数据上进行适应性微调，减少标注成本的同时提高模型的准确性和泛化性。

(7) 最近基础大模型在自然图像分割任务上取得了突破性的进展^[242]，在医学图像分割中，由于数据成像和疾病的多样性，很难构建统一的医学图像分割模型，但利用现有基础大模型根据特定临床场景构造专才模型也有巨大的实际价值^[243]。未来如何利用现有基础分割模型，结合少量特定任务的数据开发专才模型也是研究的重点。

(8) 在本研究中，除第六章方法外，其余方法均缺少开发临床应用前需要的多

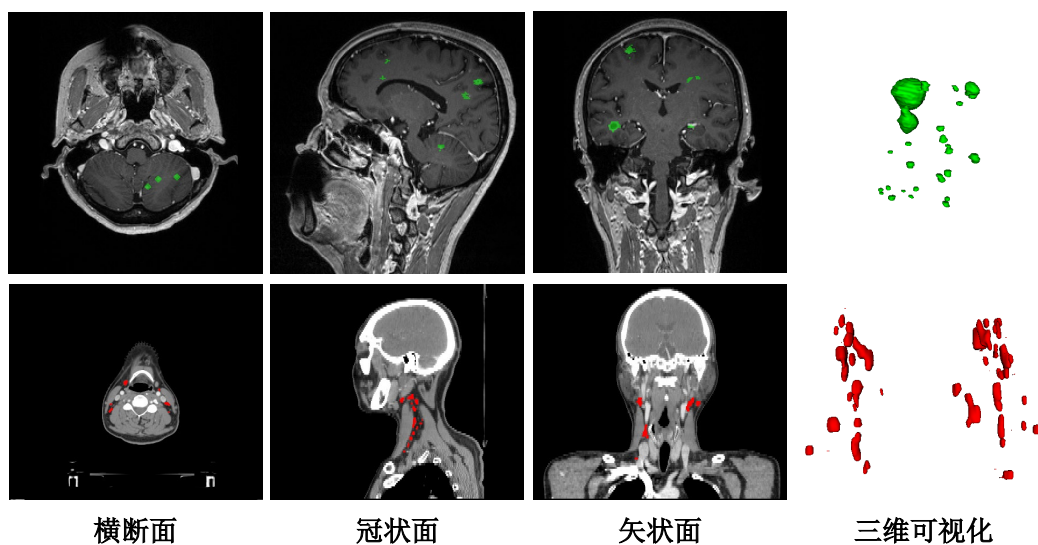


图 7-1 小病灶分割示意图，第一行为颅脑转移瘤分割任务，第二行为颈部淋巴结分割任务

中心大规模验证。此外，这些方法都只在回顾性的数据集上进行了实验，很难准确获取其各种在真实临床实际中的表现。因此，探索如何将现有深度学习方法转化为可以实际临床应用产品，进行多中心大规模的前瞻性测试也是今后研究需要重点考虑的问题。

参考文献

- [1] Budd S, Robinson E. C, Kainz B. A survey on active learning and human-in-the-loop deep learning for medical image analysis[J]. Medical Image Analysis, 2021, 71: 102062.
- [2] Boykov Y. Y, Jolly M.-P. Interactive graph cuts for optimal boundary & region segmentation of objects in ND images[C]. Proceedings of the IEEE International Conference on Computer Vision, Vancouver, British Columbia, Canada, July 7-14, 2001, 105-112.
- [3] Krizhevsky A, Sutskever I, Hinton G. E. ImageNet classification with deep convolutional neural networks[C]. Advances in Neural Information Processing Systems, Lake Tahoe, Nevada, UAS, December 3-6, 2012, 1106-1114.
- [4] Hinton G. E, Roweis S. Stochastic neighbor embedding[C]. Advances in Neural Information Processing Systems, Vancouver, British Columbia, Canada, December 9-14, 2002, 833-840.
- [5] Milletari F, Navab N, Ahmadi S.-A. V-net: Fully convolutional neural networks for volumetric medical image segmentation[C]. Proceedings of the IEEE International Conference on 3D Vision, Stanford, CA, USA, October 25-28, 2016, 565-571.
- [6] 徐宝泉, 凌彤辉. 基于级联 Vnet-S 网络的 CT 影像单一器官自动分割算法 [J]. 计算机应用, 2019, 39(8): 2420.
- [7] Tang M, Perazzi F, Djelouah A, et al. On regularized losses for weakly-supervised CNN segmentation[C]. Proceedings of the European Conference on Computer Vision, Munich, Germany, September 8-14, 2018, 507-522.
- [8] Ronneberger O, Fischer P, Brox T. U-Net: Convolutional networks for biomedical image segmentation[C]. International Conference on Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, October 5 - 9, 2015, 234-241.
- [9] Dolz J, Desrosiers C, Ayed I. B. Teach me to segment with mixed supervision: Confident students become masters[C]. International Conference on Information Processing In Medical Imaging, Virtual, June 28-June 30, 2021, 517-529.
- [10] Wu Y, Xu M, Ge Z, et al. Semi-supervised left atrium segmentation with mutual consistency training[C]. International Conference on Medical Image Computing and Computer-Assisted Intervention, Strasbourg, France, September 27 - October 1, 2021, 297-306.

- [11] Chen X, Yuan Y, Zeng G, et al. Semi-supervised semantic segmentation with cross pseudo supervision[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Virtual, June 19-25, 2021, 2613-2622.
- [12] 王云新, 王大勇, 杨怡姝, 等. 数字全息技术在生物医学成像和分析中的应用 [J]. 中国激光, 2014, 41(2): 0209002.
- [13] 唐峰, 谢晋东, 赵雷, 等. 医学成像系统显示器性能的对比研究 [J]. 中华放射学杂志, 2008, 42(2): 192-195.
- [14] 龚小竞, 孟静, 陈健桦, 等. 生物医学光声成像技术及其临床应用进展 [J]. 集成技术, 2013, 53-59.
- [15] Armato III S. G, Sensakovic W. F. Automated lung segmentation for thoracic CT: impact on computer-aided diagnosis[J]. Academic Radiology, 2004, 11(9): 1011-1021.
- [16] Antonelli M, Reinke A, Bakas S, et al. The medical segmentation decathlon[J]. Nature Communications, 2022, 13(1): 4128.
- [17] Otsu N. A threshold selection method from gray-level histograms[J]. IEEE Transactions on Systems, Man, and Cybernetics, 1979, 9(1): 62-66.
- [18] Deng Y, Manjunath B. S. Unsupervised segmentation of color-texture regions in images and video[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2001, 23(8): 800-810.
- [19] Haris K, Efstratiadis S. N, Maglaveras N, et al. Hybrid image segmentation using watersheds and fast region merging[J]. IEEE Transactions on Image Processing, 1998, 7(12): 1684-1699.
- [20] Zhu S. C, Yuille A. Region competition: Unifying snakes, region growing, and Bayes/MDL for multiband image segmentation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1996, 18(9): 884-900.
- [21] Malik J, Belongie S, Leung T, et al. Contour and texture analysis for image segmentation[J]. International Journal of Computer Vision, 2001, 43: 7-27.
- [22] Kohli P, Ladický L, Torr P. H. Robust higher order potentials for enforcing label consistency[J]. International Journal of Computer Vision, 2009, 82: 302-324.
- [23] Wang X.-Y, Wang T, Bu J. Color image segmentation using pixel wise support vector machine classification[J]. Pattern Recognition, 2011, 44(4): 777-787.
- [24] Rother C, Kolmogorov V, Blake A. GrabCut: Interactive foreground extraction using iterated graph cuts[J]. ACM Transactions on Graphics, 2004, 23(3): 309-314.

- [25] Wang G, Zuluaga M. A, Pratt R, et al. Dynamically balanced online random forests for interactive scribble-based segmentation[C]. International Conference on Medical Image Computing and Computer-Assisted Intervention, Athens, Greece, October 17-21, 2016, 352-360.
- [26] Xu C, Prince J. L. Snakes, shapes, and gradient vector flow[J]. IEEE Transactions on Image Processing, 1998, 7(3): 359-369.
- [27] Vese L. A, Chan T. F. A multiphase level set framework for image segmentation using the mumford and shah model[J]. International Journal of Computer Vision, 2002, 50: 271-293.
- [28] Viola P, Wells III W. M. Alignment by maximization of mutual information[J]. International Journal of Computer Vision, 1997, 24(2): 137-154.
- [29] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, June 27-30, 2016, 770-778.
- [30] Deng J, Dong W, Socher R, et al. ImageNet: A large-scale hierarchical image database[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Miami, Florida, USA, June 20-25, 2009, 248-255.
- [31] Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, June 7-12, 2015, 3431-3440.
- [32] Çiçek Ö, Abdulkadir A, Lienkamp S. S, et al. 3D U-Net: learning dense volumetric segmentation from sparse annotation[C]. International Conference on Medical Image Computing and Computer-Assisted Intervention, Athens, Greece, October 17-21, 2016, 424-432.
- [33] Zhou Z, Siddiquee M. M. R, Tajbakhsh N, et al. Unet++: Redesigning skip connections to exploit multiscale features in image segmentation[J]. IEEE Transactions on Medical Imaging, 2019, 39(6): 1856-1867.
- [34] Isensee F, Jaeger P. F, Kohl S. A, et al. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation[J]. Nature Methods, 2021, 18(2): 203-211.
- [35] Luo X, Song T, Wang G, et al. Scpm-net: An anchor-free 3D lung nodule detection network using sphere representation and center points matching[J]. Medical Image Analysis, 2022, 75: 102287.
- [36] Lin L, Dou Q, Jin Y.-M. Deep learning for automated contouring of primary tumor volumes by MRI for nasopharyngeal carcinoma[J]. Radiology, 2019, 291(3): 677-686.
- [37] 王晋东, 陈益强. 迁移学习导论 [M]. 电子工业出版社, 2021.

- [38] Jordan M. I, Mitchell T. M. Machine learning: Trends, perspectives, and prospects[J]. Science, 2015, 349(6245): 255-260.
- [39] 马梓博, 米悦, 张波, 等. 面向异质性医学图像处理的深度学习算法综述 [J]. 软件学报, 2023, 34: 4870-4915.
- [40] 田娟秀, 刘国才, 谷珊珊, 等. 医学图像分析深度学习方法研究与挑战 [J]. 自动化学报, 2018, 44(3): 401-424.
- [41] 曹玉红, 徐海, 刘荪傲, 等. 基于深度学习的医学影像分割研究综述 [J]. 计算机应用, 2021, 41(8): 2273-2287.
- [42] 左艳, 黄钢, 聂生东. 深度学习在医学影像智能处理中的应用与挑战 [J]. 中国图象图形学报, 2021, 26: 305-315.
- [43] Cordts M, Omran M, Ramos S, et al. The cityscapes dataset for semantic urban scene understanding[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, June 27-30, 2016, 3213-3223.
- [44] Luo X, Liao W, Xiao J, et al. WORD: A large scale dataset, benchmark and clinical applicable study for abdominal organ segmentation from ct image[J]. Medical Image Analysis, 2022, 82: 102642.
- [45] 李赞铎, 宫恩浩, 李睿, 等. 深度学习技术与医学影像——现状及未来 [J]. 中华放射学杂志, 2018, 52(5): 321-326.
- [46] 叶怀泽, 周子奇, 祁磊, 等. 基于特定领域解码的域泛化医学图像分割方法 [J]. 数据采集与处理, 2023, 38: 324-335.
- [47] Vincent O. R, Folorunso O. A descriptive algorithm for sobel image edge detection[J]. Informing Science and IT Education Journal, 2009, 40: 97-107.
- [48] Prewitt J. M. Object enhancement and extraction[J]. Picture processing and Psychopictorics, 1970, 10(1): 15-19.
- [49] Canny J. A computational approach to edge detection[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1986, 679-698.
- [50] 段锐, 管一弘. 医学图像自动多阈值分割 [J]. 计算机应用, 2008, 196-197.
- [51] 申铨京, 潘红, 陈海鹏. 基于一维 OTSU 的多阈值医学图像分割算法 [J]. 吉林大学学报: 理学版, 2016, 54(2): 344-348.
- [52] Shafarenko L, Petrou M, Kittler J. Automatic watershed segmentation of randomly textured color images[J]. IEEE Transactions on Image Processing, 1997, 6(11): 1530-1544.

- [53] Kass M, Witkin A, Terzopoulos D. Snakes: Active contour models[J]. International Journal of Computer Vision, 1988, 1(4): 321-331.
- [54] Caselles V, Kimmel R, Sapiro G. Geodesic active contours[J]. International Journal of Computer Vision, 1997, 22: 61-79.
- [55] Chan T. F, Vese L. A. Active contours without edges[J]. IEEE Transactions on Image Processing, 2001, 10(2): 266-277.
- [56] Yushkevich P. A, Piven J, Hazlett H. C, et al. User-guided 3D active contour segmentation of anatomical structures: significantly improved efficiency and reliability[J]. Neuroimage, 2006, 31(3): 1116-1128.
- [57] Liang J, McInerney T, Terzopoulos D. United snakes[J]. Medical Image Analysis, 2006, 10(2): 215-233.
- [58] Goldstein T, Li M, Yuan X. Adaptive primal-dual splitting methods for statistical learning and image processing[C]. Advances in Neural Information Processing Systems, Montreal, Quebec, Canada, December 7-12, 2015, 2089-2097.
- [59] Yuan J, Bae E, Tai X.-C. A study on continuous max-flow and min-cut approaches[C]. Proceedings of the IEEE International Conference on Computer Vision, San Francisco, CA, USA, June 13-18, 2010, 2217-2224.
- [60] Li C, Xu C, Gui C, et al. Level set evolution without re-initialization: a new variational formulation[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, San Diego, CA, USA, June 20-26, 2005, 430-436.
- [61] Cai X, Chan R, Zeng T. A two-stage image segmentation method using a convex variant of the mumford-shah model and thresholding[J]. SIAM Journal on Imaging Sciences, 2013, 6(1): 368-390.
- [62] Price B. L, Morse B, Cohen S. Geodesic graph cut for interactive image segmentation[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, San Francisco, CA, USA, June 13-18, 2010, 3161-3168.
- [63] Gulshan V, Rother C, Criminisi A, et al. Geodesic star convexity for interactive image segmentation[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, San Francisco, CA, USA, June 13-18, 2010, 3129-3136.
- [64] Wang G, Zuluaga M. A, Pratt R, et al. Slic-Seg: A minimally interactive segmentation of the placenta from sparse and motion-corrupted fetal MRI in multiple views[J]. Medical Image Analysis, 2016, 34: 137-147.

- [65] Rother C, Kolmogorov V, Blake A. GrabCut: interactive foreground extraction using iterated graph cuts[J]. TOG, 2004, 23(3): 309-314.
- [66] 郭雯, 鞠忠建, 吴青南. 基于深度学习的器官自动分割研究进展 [J]. 医疗卫生装备, 2020, 41(1): 85-94.
- [67] Liu Z, Lin Y, Cao Y, et al. Swin Transformer: Hierarchical vision transformer using shifted windows[C]. Proceedings of the IEEE International Conference on Computer Vision, Montreal, QC, Canada, October 10-17, 2021, 10012-10022.
- [68] Tang Y, Yang D, Li W, et al. Self-supervised pre-training of swin transformers for 3D medical image analysis[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, June 18-24, 2022, 20730-20740.
- [69] Zhou H.-Y, Guo J, Zhang Y, et al. nnFormer: Volumetric medical image segmentation via a 3D transformer[J]. IEEE Transactions on Image Processing, 2022, 32(1): 4036-4045.
- [70] Tajbakhsh N, Jeyaseelan L, Li Q, et al. Embracing imperfect datasets: A review of deep learning solutions for medical image segmentation[J]. Medical Image Analysis, 2020, 63: 101693.
- [71] Criminisi A, Sharp T, Blake A. GeoS: Geodesic image segmentation[C]. Proceedings of the European Conference on Computer Vision, Marseille, France, October 12-18, 2008, 99-112.
- [72] Grady L. Random walks for image segmentation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2006, 28(11): 1768-1783.
- [73] Vezhnevets V, Konouchine V. GrowCut: Interactive multi-label ND image segmentation by cellular automata[J]. Graphicon, 2005, 1(4): 150-156.
- [74] Hu Y, Soltoggio A, Lock R, et al. A fully convolutional two-stream fusion network for interactive image segmentation[J]. Neural Networks, 2019, 109: 31-42.
- [75] Wang B, Liu K. W, Prastawa K. M, et al. 4D active cut: An interactive tool for pathological anatomy modeling[C]. Proceedings of the IEEE International Symposium on Biomedical Imaging, Beijing, China, April 29 - May 2, 2014, 529-532.
- [76] Wang G, Li W, Zuluaga M. A, et al. Interactive medical image segmentation using deep learning with image-specific fine tuning[J]. IEEE Transactions on Medical Imaging, 2018, 37(7): 1562-1573.
- [77] Wang G, Zuluaga M. A, Li W, et al. DeepIGeoS: a deep interactive geodesic framework for medical image segmentation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 41(7): 1559-1572.

- [78] Shen D, Wu G, Suk H.-I. Deep learning in medical image analysis[J]. Annual Review of Biomedical Engineering, 2017, 19: 221-248.
- [79] Litjens G, Kooi T, Bejnordi B. E, et al. A survey on deep learning in medical image analysis[J]. Medical Image Analysis, 2017, 42: 60-88.
- [80] Rajchl M, Lee M. C, Oktay O, et al. DeepCut: Object segmentation from bounding box annotations using convolutional neural networks[J]. IEEE Transactions on Medical Imaging, 2016, 36(2): 674-683.
- [81] Xu N, Price B, Cohen S, et al. Deep interactive object selection[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, June 27-30, 2016, 373-381.
- [82] Maninis K.-K, Caelles S, Pont-Tuset J, et al. Deep Extreme Cut: From extreme points to object segmentation[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, June 18-22, 2018, 616-625.
- [83] Diaz-Pinto A, Mehta P, Alle S, et al. DeepEdit: deep editable learning for interactive segmentation of 3D medical images[C]. MICCAI Workshop on Data Augmentation, Labelling, and Imperfections, Singapore, September 22, 2022, 11-21.
- [84] Castrejon L, Kundu K, Urtasun R, et al. Annotating object instances with a Polygon-RNN[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, July 21-26, 2017, 5230-5238.
- [85] Acuna D, Ling H, Kar A, et al. Efficient interactive annotation of segmentation datasets with Polygon-RNN++[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, June 18-22, 2018, 859-868.
- [86] 兰红. 多阈值优化的交互式医学图像分割方法 [J]. 计算机科学, 2013, 40(9): 296-299.
- [87] Lin D, Dai J, Jia J, et al. ScribbleSup: Scribble-supervised convolutional networks for semantic segmentation[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, June 27-30, 2016, 3159-3167.
- [88] Can Y. B, Chaitanya K, Mustafa B, et al. Learning to segment medical images with scribble-supervision alone[C]. Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support, DLMIA 2018, Granada, Spain, September 20, 2018, Proceedings 4, 2018, 236-244.
- [89] Kim B, Ye J. C. Mumford-Shah loss functional for image segmentation with deep learning[J]. IEEE Transactions on Image Processing, 2019, 29: 1856-1866.

- [90] Lee H, Jeong W.-K. Scribble2Label: Scribble-supervised cell segmentation via self-generating pseudo-labels with consistency[C]. International Conference on Medical Image Computing and Computer-Assisted Intervention, Lima, Peru, October 4-8, 2020, 14-23.
- [91] Liu X, Yuan Q, Gao Y, et al. Weakly supervised segmentation of covid19 infection with scribble annotation on CT images[J]. Pattern Recognition, 2022, 122: 108341.
- [92] Valvano G, Leo A, Tsaftaris S. A. Learning to segment from scribbles using multi-scale adversarial attention gates[J]. IEEE Transactions on Medical Imaging, 2021, 40(8): 1990-2001.
- [93] Lee D.-H. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks[C]. Proceedings of the International Conference on Machine Learning, 2013, 896.
- [94] Tarvainen A, Valpola H. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results[C]. Advances in Neural Information Processing Systems, Long Beach, CA, USA, December 4-9, 2017, 1195-1204.
- [95] Bai W, Oktay O, Sinclair M, et al. Semi-supervised learning for network-based cardiac MR image segmentation[C]. International Conference on Medical Image Computing and Computer-Assisted Intervention, Quebec City, QC, Canada, September 11-13, 2017, 253-260.
- [96] Cui W, Liu Y, Li Y, et al. Semi-supervised brain lesion segmentation with an adapted mean teacher model[C]. International Conference on Information Processing In Medical Imaging, Hong Kong, China, June 2-7, 2019, 554-565.
- [97] Yu L, Wang S, Li X, et al. Uncertainty-aware self-ensembling model for semi-supervised 3D left atrium segmentation[C]. International Conference on Medical Image Computing and Computer-Assisted Intervention, Shenzhen, China, October 13-17, 2019, 605-613.
- [98] Wang Y, Zhang Y, Tian J, et al. Double-uncertainty weighted method for semi-supervised learning[C]. International Conference on Medical Image Computing and Computer-Assisted Intervention, Lima, Peru, October 4-8, 2020, 542-551.
- [99] Hang W, Feng W, Liang S, et al. Local and global structure-aware entropy regularized mean teacher model for 3D left atrium segmentation[C]. International Conference on Medical Image Computing and Computer-Assisted Intervention, Lima, Peru, October 4-8, 2020, 562-571.
- [100] Zhang Y, Yang L, Chen J, et al. Deep adversarial networks for biomedical image segmentation utilizing unannotated images[C]. International Conference on Medical Image Computing and Computer-Assisted Intervention, Quebec City, QC, Canada, September 11-13, 2017, 408-416.

- [101] Nie D, Gao Y, Wang L, et al. Asdnet: Attention based semi-supervised deep networks for medical image segmentation[C]. International Conference on Medical Image Computing and Computer-Assisted Intervention, Granada, Spain, September 16-20, 2018, 370-378.
- [102] Cohn D. A, Ghahramani Z, Jordan M. I. Active learning with statistical models[J]. Journal of Artificial Intelligence Research, 1996, 4: 129-145.
- [103] Lewis D. D, Catlett J. Heterogeneous uncertainty sampling for supervised learning[M]. Elsevier, 1994, 148-156.
- [104] Scheffer T, Wrobel S, Popov B, et al. Learning hidden markov models for information extraction actively from partially labeled text[J]. Künstliche Intell., 2002, 16(2): 17-22.
- [105] Xu Z, Yu K, Tresp V, et al. Representative sampling for text classification using support vector machines[C]. European Conference on Information Retrieval, Pisa, Italy, April 14-16, 2003, 393-407.
- [106] Huang S.-J, Jin R, Zhou Z.-H. Active learning by querying informative and representative examples[C]. Advances in Neural Information Processing Systems, Vancouver, British Columbia, Canada, December 6-9, 2010, 892-900.
- [107] Freytag A, Rodner E, Denzler J. Selecting influential examples: Active learning with expected model output changes[C]. Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, September 6-12, 2014, 562-577.
- [108] Kading C, Freytag A, Rodner E, et al. Active learning and discovery of object categories in the presence of unnameable instances[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, June 7-12, 2015, 4343-4352.
- [109] 崔文超, 王毅, 樊养余, 等. 基于 FCM 和 LBF 模型的医学图像自动分割 [J]. 计算机工程, 2012, 38(24): 200-204.
- [110] Zaitoun N. M, Aqel M. J. Survey on image segmentation techniques[J]. Procedia Computer Science, 2015, 65: 797-806.
- [111] Xu C, Pham D. L, Prince J. L. Image segmentation using deformable models[J]. Handbook of medical imaging, 2000, 2(20): 0.
- [112] Osher S, Sethian J. A. Fronts propagating with curvature-dependent speed: Algorithms based on hamilton-jacobi formulations[J]. Journal of Computational Physics, 1988, 79(1): 12-49.
- [113] Osher S, Fedkiw R, Piechor K. Level set methods and dynamic implicit surfaces[J]. Appl. Mech. Rev., 2004, 57(3): B15-B15.

- [114] Li C, Xu C, Gui C, et al. Distance regularized level set evolution and its application to image segmentation[J]. IEEE Transactions on Image Processing, 2010, 19(12): 3243-3254.
- [115] 何源, 罗予频, 胡东成. 基于梯度向量流的医学图像自动分割 [J]. 计算机应用, 2007, 27(1): 149-151.
- [116] Hu P, Shuai B, Liu J, et al. Deep level sets for salient object detection[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, July 21-26, 2017, 2300-2309.
- [117] Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, June 7-12, 2015, 1-9.
- [118] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[C]. Proceedings of the International Conference on Learning Representations, San Diego, CA, USA, May 7-9, 2014, 1-13.
- [119] He K, Zhang X, Ren S, et al. Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification[C]. Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, December 7-13, 2015, 1026-1034.
- [120] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]. Advances in Neural Information Processing Systems, Long Beach, CA, USA, December 4-9, 2017, 5998-6008.
- [121] Shamshad F, Khan S, Zamir S. W, et al. Transformers in medical imaging: A survey[J]. Medical Image Analysis, 2023, 102802.
- [122] Lin M, Chen Q, Yan S. Network in network[C]. Proceedings of the International Conference on Learning Representations, Banff, Canada, April 14-16, 2013, 1-13.
- [123] Robbins H, Monro S. A stochastic approximation method[J]. The Annals of Mathematical Statistics, 1951, 400-407.
- [124] Kingma D. P, Ba J. Adam: A method for stochastic optimization[C]. Proceedings of the International Conference on Learning Representations, San Diego, CA, USA, May 7-9, 2014, 1-13.
- [125] Kohl S, Romera-Paredes B, Meyer C, et al. A probabilistic U-Net for segmentation of ambiguous images[C]. Advances in Neural Information Processing Systems, Montréal, Canada, December 3-8, 2018, 6965-6975.
- [126] Li X, Chen H, Qi X, et al. H-denseunet: hybrid densely connected unet for liver and tumor segmentation from CT volumes[C]. IEEE Transactions on Medical Imaging, 2018, 2663-2674.

- [127] 黄晓鸣, 何富运, 唐晓虎, 等. U-net 及其变体在医学图像分割中的应用研究综述 [J]. 中国生物医学工程学报, 2022, 41(5): 567-576.
- [128] 吴玉超, 林岚, 王婧璇, et al. 基于卷积神经网络的语义分割在医学图像中的应用 [J]. 生物医学工程学报, 2020, 37: 533-540.
- [129] Chen L, Papandreou G, Kokkinos I, et al. Semantic image segmentation with deep convolutional nets and fully connected CRFs[C]. Proceedings of the International Conference on Learning Representations, San Diego, CA, USA, May 7-9, 2015, 1-13.
- [130] Chen L.-C, Papandreou G, Kokkinos I, et al. DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 40(4): 834-848.
- [131] Chen L.-C, Zhu Y, Papandreou G, et al. Encoder-decoder with atrous separable convolution for semantic image segmentation[C]. Proceedings of the European Conference on Computer Vision, Munich, Germany, September 8-14, 2018, 801-818.
- [132] Zhao H, Shi J, Qi X, et al. Pyramid scene parsing network[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, July 21-26, 2017, 2881-2890.
- [133] Lee C.-Y, Xie S, Gallagher P, et al. Deeply-supervised nets[C]. Artificial Intelligence and Statistics, San Diego, California, USA, May 9-12, 2015, 562-570.
- [134] Kamnitsas K, Ledig C, Newcombe V. F, et al. Efficient multi-scale 3D CNN with fully connected crf for accurate brain lesion segmentation[J]. Medical Image Analysis, 2017, 36: 61-78.
- [135] Kamnitsas K, Ferrante E, Parisot S, et al. DeepMedic for brain tumor segmentation[C]. International workshop on Brainlesion: Glioma, multiple sclerosis, stroke and traumatic brain injuries, Athens, Greece, October 17, 2016, 138-149.
- [136] Hu J, Shen L, Sun G. Squeeze-and-excitation networks[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, June 18-22, 2018, 7132-7141.
- [137] Woo S, Park J, Lee J.-Y, et al. CBAM: Convolutional block attention module[C]. Proceedings of the European Conference on Computer Vision, Munich, Germany, September 8-14, 2018, 3-19.
- [138] Fu J, Liu J, Tian H, et al. Dual attention network for scene segmentation[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, June 16-20, 2019, 3146-3154.

- [139] Schlemper J, Oktay O, Schaap M, et al. Attention gated networks: Learning to leverage salient regions in medical images[J]. *Medical Image Analysis*, 2019, 53: 197-207.
- [140] Wang Y, Wei X, Liu F, et al. Deep distance transform for tubular structure segmentation in CT scans[C]. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Seattle, WA, USA, June 13-19, 2020, 3833-3842.
- [141] Chen J, Lu Y, Yu Q, et al. TransUnet: Transformers make strong encoders for medical image segmentation[J]. *arXiv preprint arXiv:2102.04306*, 2021.
- [142] Xie E, Wang W, Yu Z, et al. SegFormer: Simple and efficient design for semantic segmentation with transformers[C]. *Advances in Neural Information Processing Systems*, Virtual, December 6-14, 2021, 12077-12090.
- [143] Zhou B, Zhao H, Puig X, et al. Semantic understanding of scenes through the ade20k dataset[J]. *International Journal of Computer Vision*, 2019, 127: 302-321.
- [144] Sohn K, Berthelot D, Carlini N, et al. FixMatch: Simplifying semi-supervised learning with consistency and confidence[C]. *Advances in Neural Information Processing Systems*, Virtual, December 6-12, 2020, 596-608.
- [145] Berthelot D, Carlini N, Goodfellow I, et al. MixMatch: A holistic approach to semi-supervised learning[C]. *Advances in Neural Information Processing Systems*, Vancouver, BC, Canada, December 8-14, 2019, 5049-5059.
- [146] Xie Q, Luong M.-T, Hovy E, et al. Self-training with noisy student improves ImageNet classification[C]. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Seattle, WA, USA, June 13-19, 2020, 10687-10698.
- [147] Huang Z, Wang X, Wang J, et al. Weakly-supervised semantic segmentation network with deep seeded region growing[C]. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, June 18-22, 2018, 7014-7023.
- [148] Felzenszwalb P. F, Huttenlocher D. P. Efficient graph-based image segmentation[J]. *International Journal of Computer Vision*, 2004, 59: 167-181.
- [149] Tang M, Djelouah A, Perazzi F, et al. Normalized cut loss for weakly-supervised CNN segmentation[C]. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, June 18-22, 2018, 1818-1827.
- [150] 刘建伟, 刘媛, 罗雄麟. 半监督学习方法 [J]. *计算机学报*, 2015, 38(8): 1592-1617.

- [151] Zou Y, Zhang Z, Zhang H, et al. PseudoSeg: Designing pseudo labels for semantic segmentation[C]. Proceedings of the International Conference on Learning Representations, Virtual, May 3-7, 2020, 1-13.
- [152] Gal Y, Ghahramani Z. Dropout as a bayesian approximation: Representing model uncertainty in deep learning[C]. Proceedings of the International Conference on Machine Learning, New York City, NY, USA, June 19-24, 2016, 1050-1059.
- [153] Luo X, Wang G, Liao W, et al. Semi-supervised medical image segmentation via uncertainty rectified pyramid consistency[J]. Medical Image Analysis, 2022, 80: 102517.
- [154] Wang G, Zhai S, Lasio G, et al. Semi-supervised segmentation of radiation-induced pulmonary fibrosis from lung ct scans with multi-scale guided dense attention[J]. IEEE transactions on medical imaging, 2021, 41(3): 531-542.
- [155] Li X, Yu L, Chen H, et al. Transformation-consistent self-ensembling model for semisupervised medical image segmentation[J]. IEEE Transactions on Neural Networks and Learning Systems, 2020, 32(2): 523-534.
- [156] Peng J, Estrada G, Pedersoli M, et al. Deep co-training for semi-supervised image segmentation[J]. Pattern Recognition, 2020, 107: 107269.
- [157] Xia Y, Yang D, Yu Z, et al. Uncertainty-aware multi-view co-training for semi-supervised medical image segmentation and domain adaptation[J]. Medical Image Analysis, 2020, 65: 101766.
- [158] Ouali Y, Hudelot C, Tami M. Semi-supervised semantic segmentation with cross-consistency training[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, June 13-19, 2020, 12674-12684.
- [159] Zhou Z, Shin J, Zhang L, et al. Fine-tuning convolutional neural networks for biomedical image analysis: actively and incrementally[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, July 21-26, 2017, 7340-7351.
- [160] Zhu J, Wang H, Yao T, et al. Active learning with sampling by uncertainty and density for word sense disambiguation and text classification[C]. International Conference on Computational Linguistics, Manchester, UK, August 18-22, 2008, 1137-1144.
- [161] Ren P, Xiao Y, Chang X, et al. A survey of deep active learning[J]. ACM Computing Surveys, 2021, 54(9): 1-40.
- [162] Settles B. From theories to queries: Active learning in practice[C]. International Artificial Intelligence and Statistics Conference, Sardinia, Italy, May 16, 2011, 1-18.

- [163] Shannon C. E. A mathematical theory of communication[J]. The Bell System Technical Journal, 1948, 27(3): 379-423.
- [164] Huang J. Z, Ng M. K, Rong H, et al. Automated variable weighting in K-Means type clustering[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2005, 27(5): 657-668.
- [165] Margatina K, Vernikos G, Barrault L, et al. Active learning by acquiring contrastive examples[C]. Proceedings of the Empirical Methods in Natural Language Processing, Virtual, November 7-11, Online and Punta Cana, Dominican Republic, 2021, 650-663.
- [166] Menze B. H, Jakab A, Bauer S, et al. The multimodal brain tumor image segmentation benchmark (brats)[J]. IEEE Transactions on Medical Imaging, 2014, 34(10): 1993-2024.
- [167] Liao X, Li W, Xu Q, et al. Iteratively-refined interactive 3D medical image segmentation with multi-agent reinforcement learning[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, June 13-19, 2020, 9394-9402.
- [168] Zhou B, Chen L, Wang Z. Interactive deep editing framework for medical image segmentation[C]. International Conference on Medical Image Computing and Computer-Assisted Intervention, Shenzhen, China, October 13-17, 2019, 329-337.
- [169] Xiong Z, Xia Q, Hu Z, et al. A global benchmark of algorithms for segmenting the left atrium from late gadolinium-enhanced cardiac magnetic resonance imaging[J]. Medical Image Analysis, 2021, 67: 101832.
- [170] Luo X, Chen J, Song T, et al. Semi-supervised medical image segmentation through dual-task consistency[C]. Proceedings of the AAAI Conference on Artificial Intelligence, Virtual, February 2-9, 2021, 8801-8809.
- [171] Li S, Zhang C, He X. Shape-aware semi-supervised 3D semantic segmentation for medical images[C]. International Conference on Medical Image Computing and Computer-Assisted Intervention, Lima, Peru, October 4-8, 2020, 552-561.
- [172] Roth H. R, Lu L, Farag A, et al. DeepOrgan: Multi-level deep convolutional networks for automated pancreas segmentation[C]. International Conference on Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, October 5-9, 2015, 556-564.
- [173] Zhou Y, Wang Y, Tang P, et al. Semi-supervised 3D abdominal multi-organ segmentation via deep multi-planar co-training[C]. Proceedings of the IEEE Winter Conference on Applications of Computer Vision, Waikoloa Village, HI, USA, January 7-11, 2019, 121-140.

- [174] Bernard O, Lalande A, Zotti C, et al. Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: is the problem solved?[J]. IEEE Transactions on Medical Imaging, 2018, 37(11): 2514-2525.
- [175] Luo X, Hu M, Song T, et al. Semi-supervised medical image segmentation via cross teaching between CNN and transformer[C]. Proceedings of the International Conference on Medical Imaging with Deep Learning, Zurich, Switzerland, July 6-8, 2022, 820-833.
- [176] 傅玉川, 余行. 医学影像自动分割技术在放射治疗中的应用及发展趋势 [J]. 中国医疗器械杂志, 2020, 44: 420-424.
- [177] Zhao F, Xie X. An overview of interactive medical image segmentation[J]. Annals of the British Machine Vision Association, 2013, 2013(7): 1-22.
- [178] Masood S, Sharif M, Masood A, et al. A survey on medical image segmentation[J]. Current Medical Imaging Reviews, 2015, 11(1): 3-14.
- [179] 刘霖雨, 龚敬, 聂生东. 利用测地线距离直方图分割血管粘连型肺结节的算法 [J]. 计算机辅助设计与图形学学报, 2017, 29(3): 493-500.
- [180] Li Z, Chen Q, Koltun V. Interactive image segmentation with latent diversity[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, June 18-22, 2018, 577-585.
- [181] Benenson R, Popov S, Ferrari V. Large-scale interactive object segmentation with human annotators[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, June 16-20, 2019, 11700-11709.
- [182] Liew J. H, Cohen S, Price B, et al. MultiSeg: Semantically meaningful, scale-diverse segmentations from minimal user input[C]. Proceedings of the IEEE International Conference on Computer Vision, Seoul, Korea (South), October 27 - November 2, 2019, 662-670.
- [183] Wang Z, Acuna D, Ling H, et al. Object instance annotation with deep extreme level set evolution[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, June 16-20, 2019, 7500-7508.
- [184] Khan S, Shahin A. H, Villafruela J, et al. Extreme points derived confidence map as a cue for class-agnostic interactive segmentation using deep neural network[C]. International Conference on Medical Image Computing and Computer-Assisted Intervention, Shenzhen, China, October 13-17, 2019, 66-73.
- [185] Bai X, Sapiro G. Geodesic matting: A framework for fast interactive image and video segmentation and matting[J]. International Journal of Computer Vision, 2009, 82(2): 113-132.

-
- [186] Papadopoulos D. P, Uijlings J. R, Keller F, et al. Extreme clicking for efficient object annotation[C]. Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, October 22-29, 2017, 4930-4939.
- [187] Harris C. G, Stephens M. A combined corner and edge detector[J]. Alvey Vision Conference, 1988, 15(50): 10-5244.
- [188] Lafferty J, McCallum A, Pereira F. C. Conditional random fields: Probabilistic models for segmenting and labeling sequence data[C]. Proceedings of the International Conference on Machine Learning, Williams College, Williamstown, MA, USA, June 28-July 1, 2001, 282-289.
- [189] Paszke A, Gross S, Massa F, et al. PyTorch: An imperative style, high-performance deep learning library[C]. Advances in Neural Information Processing Systems, Vancouver, BC, Canada, December 8-14, 2019, 8026-8037.
- [190] Xu N, Price B. L, Cohen S, et al. Deep GrabCut for object selection[C]. Proceedings of the British Machine Vision Conference, London, UK, September 4-7, 2017, 1-12.
- [191] Wang G, Aertsen M, Deprest J, et al. Uncertainty-guided efficient interactive refinement of fetal brain segmentation from stacks of MRI slices[C]. International Conference on Medical Image Computing and Computer-Assisted Intervention, Lima, Peru, October 4-8, 2020, 279-288.
- [192] Tajbakhsh N, Shin J. Y, Gurudu S. R, et al. Convolutional neural networks for medical image analysis: Full training or fine tuning?[J]. IEEE Transactions on Medical Imaging, 2016, 35(5): 1299-1312.
- [193] Sourati J, Gholipour A, Dy J. G, et al. Intelligent labeling based on fisher information for medical image segmentation using deep learning[J]. IEEE Transactions on Medical Imaging, 2019, 38(11): 2642-2653.
- [194] Rupprecht C, Laina I, Navab N, et al. Guide me: Interacting with deep networks[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, June 18-22, 2018, 8551-8561.
- [195] Zhou X, Zhuo J, Krahenbuhl P. Bottom-up object detection by grouping extreme and center points[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, June 16-20, 2019, 850-859.
- [196] Song G, Myeong H, Mu Lee K. SeedNet: Automatic seed generation with deep reinforcement learning for robust interactive segmentation[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, June 18-22, 2018, 1760-1768.

- [197] Top A, Hamarneh G, Abugharbieh R. Active learning for interactive 3D image segmentation[C]. International Conference on Medical Image Computing and Computer-Assisted Intervention, Toronto, Canada, September 18-22, 2011, 603-610.
- [198] Wang X, Gao J, Long M, et al. Self-tuning for data-efficient deep learning[C]. Proceedings of the International Conference on Machine Learning, Virtual, July 18-24, 2021, 10738-10748.
- [199] Huo X, Xie L, He J, et al. ATSO: Asynchronous teacher-student optimization for semi-supervised image segmentation[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Virtual, June 19-25, 2021, 1235-1244.
- [200] Zhang H, Cisse M, Dauphin Y. N, et al. Mixup: Beyond empirical risk minimization[C]. Proceedings of the International Conference on Learning Representations, Vancouver, BC, Canada, April 30 - May 3, 2017, 1-13.
- [201] Grandvalet Y, Bengio Y. Semi-supervised learning by entropy minimization[C]. Advances in Neural Information Processing Systems, Vancouver, British Columbia, Canada, December 13-18, 2005, 529-536.
- [202] Dorent R, Joutard S, Shapey J, et al. Inter extreme points geodesics for end-to-end weakly supervised image segmentation[C]. International Conference on Medical Image Computing and Computer-Assisted Intervention, Strasbourg, France, September 27 - October 1, 2021, 615-624.
- [203] 刘少鹏, 洪佳明, 梁杰鹏, 等. 面向医学图像分割的半监督条件生成对抗网络 [J]. 软件学报, 2020, 31(8): 2588-2602.
- [204] Vu T.-H, Jain H, Bucher M, et al. AdvEnt: Adversarial entropy minimization for domain adaptation in semantic segmentation[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, June 16-20, 2019, 2517-2526.
- [205] Luo X, Wang G, Song T, et al. MIDeepSeg: Minimally interactive segmentation of unseen objects from medical images using deep learning[J]. Medical Image Analysis, 2021, 72: 102102.
- [206] Luo X, Liao W, Chen J, et al. Efficient semi-supervised gross target volume of nasopharyngeal carcinoma segmentation via uncertainty rectified pyramid consistency[C]. International Conference on Medical Image Computing and Computer-Assisted Intervention, Strasbourg, France, September 27 - October 1, 2021, 318-329.
- [207] Ma J, Wei Z, Zhang Y, et al. How distance transform maps boost segmentation CNNs: An empirical study[C]. Proceedings of the International Conference on Medical Imaging with Deep Learning, Montréal, QC, Canada, July 6-8, 2020, 479-492.

- [208] Ma J, He J, Yang X. Learning geodesic active contours for embedding object global information in segmentation CNNs[J]. IEEE Transactions on Medical Imaging, 2020, 40(1): 93-104.
- [209] Xue Y, Tang H, Qiao Z, et al. Shape-aware organ segmentation by predicting signed distance maps[C]. Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, February 7-12, 2020, 12565-12572.
- [210] Zhen M, Wang J, Zhou L, et al. Joint semantic segmentation and boundary detection using iterative pyramid contexts[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, June 13-19, 2020, 13666-13675.
- [211] Cheng B, Xiao B, Wang J, et al. Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, June 13-19, 2020, 5386-5395.
- [212] Simonyan K, Zisserman A. Two-stream convolutional networks for action recognition in videos[C]. Advances in Neural Information Processing Systems, Montreal, Quebec, Canada, December 8-13, 2014, 568-576.
- [213] Zamir A. R, Sax A, Shen W, et al. Taskonomy: Disentangling task transfer learning[C]. Proceedings of the International Joint Conference on Artificial Intelligence, Macao, China, August 10-16, 2018, 3712-3722.
- [214] Zamir A. R, Sax A, Cheerla N, et al. Robust learning through cross-task consistency[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, June 13-19, 2020, 11197-11206.
- [215] Lever J, Krzywinski M, Altman N. Points of significance: model selection and overfitting[J]. Nature Methods, 2016, 13(9): 703-705.
- [216] MacQueen J. Some methods for classification and analysis of multivariate observations[J]. Symposium on mathematical statistics and probability, 1967, 1(14): 281-297.
- [217] Wang G, Luo X, Gu R, et al. PyMIC: A deep learning toolkit for annotation-efficient medical image segmentation[J]. Computer Methods and Programs in Biomedicine, 2023, 231: 107398.
- [218] Kushibar K, Valverde S, Gonzalez-Villa S, et al. Supervised domain adaptation for automatic sub-cortical brain structure segmentation with minimal user interaction[J]. Scientific reports, 2019, 9(1): 6742.
- [219] Tsai Y.-H, Hung W.-C. Learning to adapt structured output space for semantic segmentation[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, June 18-22, 2018, 7472-7481.

- [220] Fleuret F. Uncertainty reduction for model adaptation in semantic segmentation[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, virtual, June 19-25, 2021, 9613-9623.
- [221] Wang D, Shelhamer E, Liu S, et al. TENT: Fully test-time adaptation by entropy minimization[C]. Proceedings of the International Conference on Learning Representations, Virtual, May 3-7, 2021, 1-13.
- [222] Chen C, Liu Q, Jin Y, et al. Source-free domain adaptive fundus image segmentation with denoised pseudo-labeling[C]. International Conference on Medical Image Computing and Computer-Assisted Intervention, Strasbourg, France, September 27 - October 1, 2021, 225-235.
- [223] Yang C, Guo X, Chen Z, et al. Source free domain adaptation for medical image segmentation with fourier style mining[J]. Medical Image Analysis, 2022, 79: 102457.
- [224] Bateson M, Kervadec H, Dolz J, et al. Source-free domain adaptation for image segmentation[J]. Medical Image Analysis, 2022, 82: 102617.
- [225] Hu S, Liao Z, Xia Y. ProSFDA: Prompt learning based source-free domain adaptation for medical image segmentation[J]. arXiv preprint arXiv:2211.11514, 2022.
- [226] Tang L, Li K, He C, et al. Source-free domain adaptive fundus image segmentation with class-balanced mean teacher[C]. International Conference on Medical Image Computing and Computer-Assisted Intervention, Vancouver, BC, Canada, October 8-12, 2023, 684-694.
- [227] Huai Z, Ding X, Li Y, et al. Context-aware pseudo-label refinement for source-free domain adaptive fundus image segmentation[C]. International Conference on Medical Image Computing and Computer-Assisted Intervention, Vancouver, BC, Canada, October 8-12, 2023, 618-628.
- [228] Su J.-C, Tsai Y.-H, Sohn K, et al. Active adversarial domain adaptation[C]. Proceedings of the IEEE Winter Conference on Applications of Computer Vision, Snowmass Village, CO, USA, March 1-5, 2020, 739-748.
- [229] Wang F, Han Z, Zhang Z, et al. MHPL: Minimum happy points learning for active source free domain adaptation[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, June 17-24, 2023, 20008-20018.
- [230] 江贵平, 秦文健, 周寿军, 等. 医学图像分割及其发展现状 [J]. 计算机学报, 2015, 38(6): 1222-1242.
- [231] Xia P, Fu K. K. Comparison of treatment plans involving intensity-modulated radiotherapy for nasopharyngeal carcinoma[J]. International Journal of Radiation Oncology* Biology* Physics, 2000, 48(2): 329-337.

- [232] Ng W. T, Chow J. C. Current radiotherapy considerations for nasopharyngeal carcinoma[J]. *Cancers*, 2022, 14(23): 5773.
- [233] Li Y, Dan T, Li H, et al. NPCNet: jointly segment primary nasopharyngeal carcinoma tumors and metastatic lymph nodes in mr images[J]. *IEEE Transactions on Medical Imaging*, 2022, 41(7): 1639-1650.
- [234] Zhou J, Chan K. L, Xu P, et al. Nasopharyngeal carcinoma lesion segmentation from mr images by support vector machine[C]. *Proceedings of the IEEE International Symposium on Biomedical Imaging*, Arlington, VA, USA, April 6-9, 2006, 1364-1367.
- [235] Huang W, Chan K. L, Zhou J. Region-based nasopharyngeal carcinoma lesion segmentation from MRI using clustering-and classification-based methods with learning[J]. *Journal of digital imaging*, 2013, 26: 472-482.
- [236] Men K, Chen X, Zhang Y, et al. Deep deconvolutional neural network for target segmentation of nasopharyngeal cancer in planning computed tomography images[J]. *Frontiers in Oncology*, 2017, 7: 315.
- [237] Chen H, Qi Y, Yin Y. MMFNet: A multi-modality MRI fusion network for segmentation of nasopharyngeal carcinoma[J]. *Neurocomputing*, 2020, 394: 27-40.
- [238] Guo Y, Yang Q, Hu W, et al. Automatic segmentation of nasopharyngeal carcinoma on mr images: A single-institution experience[J]. *International Journal of Radiation Oncology* Biology* Physics*, 2020, 108(3): e776.
- [239] Shirikhorshidi A. S, Aghabozorgi S, Wah T. Y. A comparison study on similarity and dissimilarity measures in clustering continuous data[J]. *PloS one*, 2015, 10(12): e0144059.
- [240] Giraud C. Introduction to high-dimensional statistics[M]. CRC Press, 2021.
- [241] Zhu J, Li Y, Hu Y, et al. Rubik's Cube+: A self-supervised feature learning framework for 3D medical image analysis[J]. *Medical Image Analysis*, 2020, 101746.
- [242] Kirillov A, Mintun E, Ravi N, et al. Segment anything[C]. *Proceedings of the IEEE International Conference on Computer Vision*, Paris, France, October 1-6, 2023, 4015-4026.
- [243] Zhang S, Metaxas D. On the challenges and perspectives of foundation models for medical image analysis[J]. *Medical Image Analysis*, 2023, 102996.