中图法分类号:TP391.4　　**文献标识码**:A　　**文章编号**:1006-8961(2025)05-1303-15

论文引用格式:Huo Y R, Feng J, Liu N, Shi Y C and Yin M Y. 2025. Ultrasound image segmentation using SAM combined with counterfactual prompt and cascaded decoder. Journal of Image and Graphics, 30(5):1303-1317(霍一儒, 封筠, 刘娜, 史屹琛, 殷梦莹. 2025. 结合反事实提示与级联解码SAM的超声图像分割. 中国图象图形学报, 30(5):1303-1317)[DOI:10.11834/jig.240447]

结合反事实提示与级联解码SAM的超声图像分割

霍一儒[1,2],封筠[1,2*],刘娜[1,2],史屹琛[3],殷梦莹[4]

1. 石家庄铁道大学信息科学与技术学院,石家庄　050043; 2. 河北省电磁环境效应与信息处理重点实验室,石家庄　050043; 3. 上海交通大学电子信息与电气工程学院,上海　200240; 4. 河北医科大学第一医院智慧医院建设部,石家庄　050023

摘　要:**目的**　分割一切模型(segment anything model,SAM)在自然图像分割领域已取得显著成就,但应用于医学成像尤其是涉及对比度低、边界模糊和形状复杂的超声图像时,分割过程往往需要人工干预,并且会出现分割性能下降情况。针对上述问题,提出一种结合反事实提示与级联解码SAM的改进方法(SAM combined wihongth counterfactual prompt and cascaded decoder,SAMCD)。**方法**　SAMCD在SAM的基础上增加旁路卷积神经网络(convolutional renual network,CNN)图像编码器、跨分支交互适配器、提示生成器和级联解码器。通过使用旁路CNN图像编码器以及跨分支交互适配器,补充ViT(vision Transformer)编码器缺乏的局部信息,以提高模型对细节的捕捉能力;引入反事实干预机制,通过生成反事实提示,迫使模型专注于事实提示生成,提高模型分割精度;采用级联解码器获得丰富的边缘信息,即先利用SAM的原始解码器创建先验掩码,再使用加入边界注意力的Transformer解码器和像素解码器;在训练模型时采用两阶段的训练策略,即交互分割模型训练阶段和自动分割模型训练阶段。**结果**　在TN3K(thyroid nodule 3K)和BUSI(breast ultrasound image)数据集上进行实验,SAMCD的DSC(Dice similarity coefficient)值分别达到83.66%和84.29%,较SAMCT提升0.73%和0.90%,且较对比的SAM及其变体模型更为轻量化;相较于9种先进方法,SAMCD在DSC、mIoU(mean intersection over union)、HD(Hausdorff distance)、敏感性和特异性指标上均达到最优。消融实验和可视化分析表明提出的SAMCD方法具有明显的提升效果。**结论**　本文提出的超声医学图像分割SAMCD方法在充分利用SAM强大的特征表达能力的基础上,通过对编码器、提示生成器、解码器和训练策略的改进,能够精准地捕获超声图像中的复杂局部细节和小目标,提高超声医学图像自动分割效果。

关键词:超声图像分割;分割一切模型(SAM);级联解码;反事实提示生成;跨分支交互适配器

Ultrasound image segmentation using SAM combined with counterfactual prompt and cascaded decoder

Huo Yiru[1,2], Feng Jun[1,2*], Liu Na[1,2], Shi Yichen[3], Yin Mengying[4]

1. School of Information Science and Technology, Shijiazhuang Tiedao University, Shijiazhuang　050043, China;
2. Hebei Key Laboratory of Electromagnetic Environmental Effects and Information Processing, Shijiazhuang　050043, China;
3. School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, Shanghai　200240, China;
4. Department of Smart Hospital Construction, The First Hospital of Hebei Medical University, Shijiazhuang　050023, China

Abstract: **Objective** Ultrasound imaging plays a crucial role in medical diagnosis due to its convenience, non-invasive nature, and cost-effectiveness, making it an essential tool in clinical settings. However, accurately localizing and extract-

收稿日期:2024-08-23;**修回日期**:2024-10-16;**预印本日期**:2024-10-23

*通信作者:封筠　fengjun@stdu.edu.cn

基金项目:国家自然科学基金项目(61972267)

Supported by:National Natural Science Foundation of China(61972267)

ing detailed features from ultrasound images, especially when dealing with complex pathological boundaries such as nodules and cysts, remains a major challenge. While traditional convolutional neural networks (CNNs) excel at feature extraction through convolutional layers, their limited receptive fields often result in the loss of crucial global information. Conversely, Transformer-based models are proficient at capturing global features through self-attention mechanisms but tend to struggle with capturing fine local details effectively. Additionally, their high computational requirements limit their use in real-time medical applications. The segmentation any model (SAM) has recently demonstrated success in natural image segmentation. However, its performance notably declines when applied to medical images, particularly ultrasound, often necessitating manual intervention. This decline is primarily due to the fact that SAM is trained exclusively on natural images, which differ substantially in domain distribution from medical images. An enhanced SAM model (i. e. , SAM combined with counterfactual prompt and cascaded decoder (SAMCD)) is proposed to address this limitation. **Method** SAMCD enhances the existing SAM framework by incorporating a Bypass CNN image encoder, a simple cross-branch interaction adapter (SCIA), a counterfactual intervention prompt generator, and a cascaded decoder. First, a Bypass CNN encoder and a novel module named SCIA are used. The limited local information of the ViT encoder is compensated by integrating the Bypass CNN encoder with the SCIA module, thereby enhancing the capability of the model to capture fine details. Next, a counterfactual intervention mechanism based on causal learning is introduced to adapt to the prompts produced by the prompt generator and optimize its output. This mechanism encourages the model to focus on generating factual prompts, strengthening the learning capability of the prompt generator, improving segmentation precision, and reducing dependency on high-quality prompts. Furthermore, a cascaded decoder is incorporated to capture rich edge information. The original SAM decoder is used to create a prior mask, followed by an edge-attention-enhanced Transformer decoder and a pixel-level decoder. This multistage decoding process enables the model to better capture and refine rich edge information, leading to highly accurate segmentation results. Finally, a two-stage training strategy is employed to enhance the segmentation performance of the model and accelerate convergence. The first stage focuses on training the interactive segmentation model, while the second stage concentrates on training the automatic segmentation model that incorporates a prompt generator. In the experiments, the hardware platform is NVIDIA GeForce RTX 3090, the programming language is Python 3. 9, and the deep learning framework is PyTorch. The network is trained under a batch size of 4, a learning rate of 0. 000 1, and a number of training rounds of 200, and the Adam optimizer is selected. SAMCD is initialized with SAM weights before training, and the images are scaled to 256 × 256 pixels using bilinear interpolation during training. **Result** Experiments were conducted on the TN3K and BUSI datasets to evaluate the performance of the SAMCD model, using a range of metrics including Dice similarity coefficient (DSC), mean intersection over union (mIoU), Hausdorff distance (HD), accuracy (Acc), sensitivity (Sen), and specificity (Spe). Notably, lower HD values indicate better segmentation performance, while higher values for metrics such as DSC and mIoU (ranging from 0 to 1) indicate better performance. In these evaluations, the SAMCD model achieved a DSC score of 83. 66% on the thyroid nodule 3K (TN3K) dataset and 84. 29% on the breast ultrasound image (BUSI) dataset, surpassing the performance of the original SAM, MedSAM, SAMed, and SAMCT in almost all metrics. Compared to SAMCT, the SAMCD model shows improvements of 0. 91% and 0. 16% on mIoU and Acc, respectively, on the TN3K dataset. Furthermore, SAMCD outperforms other SAM-related comparison models by an average of 20. 43% and 12. 91% in performance. When compared to non-SAM approaches, SAMCD achieves 4. 65%, 3. 29%, 13. 58%, 5. 16%, and 2. 22% higher DSC values than U-Net, CE-Net, SwinUnet, TransFuse, and TransUNet, respectively, on the TN3K dataset. In terms of Acc, SAMCD outperforms TransFuse and TransUNet by 0. 79% and 0. 29%, respectively, while also demonstrating an average improvement of 4. 95% in Sen and 0. 46% in Spe over the five non-SAM methods. Additionally, SAMCD requires fewer training parameters and consumes less computational resources compared to SAM-related models. Ablation experiments and visual analyses further confirm the substantial performance improvements provided by the SAMCD method. **Conclusion** SAMCD leverages the strong feature extraction capabilities of SAM by enhancing its encoder, prompt generator, decoder, and training strategy. These improvements enable SAMCD to accurately capture complex local details and small targets in ultrasound images, thereby substantially improving the automatic segmentation performance for ultrasound medical imaging.

Key words: ultrasound image segmentation; segment anything model (SAM); cascaded decoding; counterfactual prompt generation; cross-branch interaction adapter

0 引 言

医学图像分割在临床诊断、疾病监测和手术规划中起着至关重要的作用。精准的医学图像分割方法能够辅助医生进行病灶定位和病变分析，显著提升诊疗效率（Noble和Boukerroui，2006）。超声作为一种广泛应用于医学领域的成像技术，因其实时性、无辐射和低成本等优点，已成为临床诊断中不可或缺的工具（Wang等，2021）。相较于计算机断层扫描（computed tomography，CT）（刘忠利 等，2018）和磁共振成像（magnetic resonance imaging，MRI）（夏峰 等，2022）等其他医学成像技术，超声图像在成像质量上存在一定的局限性，特别是图像的低对比度、边界模糊以及组织的异质性等特征，这使得自动超声图像分割变得尤为困难（于典 等，2023）。因此，研究如何提高超声医学图像分割的准确性，克服成像质量带来的挑战，显得尤为重要。

目前，基于深度学习的方法（Chen等，2020）已广泛应用于超声医学图像分割，大致可分为基于卷积神经网络（convolutional neural network，CNN）（Dhruv和Naskar，2020）和基于Transformer（Xiao等，2023）的两类方法。U-Net（Ronneberger等，2015）是CNN方法中的开创性工作，具有出色的保留细节的能力。受U-Net的启发，研究人员提出一系列改进，如CE-Net（context encoder network）（Gu等，2019）通过引入空洞卷积和残差块来解决U-Net空间信息的丢失问题，以提高分割的精度。随着Transformer在计算机视觉领域的成功应用（陈洛轩 等，2023），基于Transformer的模型被引入医学图像分割。例如，TransUNet（Chen等，2021）在U-Net模型的基础上引入混合编码器，将CNN和Transformer结合起来，成为最具代表性的方法之一；SwinUnet（Cao等，2023）与TransUNet不同，其利用Swin Transformer块（Han等，2023）从输入图像中提取分层特征；TransFuse（Zhang等，2021）则是通过采用分支并行结构，结合CNN和Transformer捕获局部细节，提高全局上下文的建模效率。这些模型的共同特点是构建一个具有跳跃连接的U形框架，在捕捉图像的全局信息同时保持图像的局部细节，从而提高分割精度。

2022年以来，基础模型（foundation models，FM）的出现引发智能模型开发的范式转变，预训练的大模型适应各种下游任务的方法变得越来越流行。分割一切模型（segment anything model，SAM）（Kirillov等，2023）因其交互功能而广受好评，在自然图像分割领域取得令人瞩目的成果（王淼 等，2024）。与经典的深度学习分割模型不同，SAM可以处理跨域的各种零样本任务，在面临不同的特定任务时，其不需要针对特定数据集重新训练模型，而是可以根据不同的提示（如点、框、文本等）输出相应的分割结果。但是目前一些研究表明（Shi等，2023；Wu等，2023；Cheng等，2023），SAM在医学图像（如超声图像）分割任务的表现不佳，主要原因是SAM仅在自然图像上进行训练，而医学图像和自然图像之间存在较大的域分布差异。

一个直接的解决方法是用医学图像对SAM进行微调。例如，MedSAM（Ma等，2024）在自行收集的多模态数据集上对SAM进行训练，用于通用医学图像分割。然而，所使用的训练数据集不够大，且存在模态不平衡的问题，MedSAM在超声图像上性能受到限制。最近有研究采用参数高效微调（parameter efficient fine tuning，PEFT）技术（Fu等，2023），并在医学成像任务上有着不错的表现。例如，Zhang和Liu（2023）提出SAMed（segment anything model for medical），引入低秩自适应（low-rank adaptation，LoRA）策略，对SAM图像编码器进行微调，以适应医学图像分割任务，在降低训练开销的同时有着不错的分割结果。Lin等人（2024）提出SAMCT，结合CNN编码器和跨分支交互模块（cross-branch interaction module，CIM）来微调SAM，但该模块的复杂性很高，这增加了微调的负担，且其提示生成器容易受混杂因素影响，SAMCT的解码器未能完全融合CNN编码器捕获的超声图像细微病变的线索，不足以完全将超声图像的细节特征进行恢复。

以上研究成果关注到不同模态医学图像的特点，分别提出针对性的改进方法，获得较好的效果。但是，由于超声图像存在对比度低、噪声大和边缘模糊等问题，因此现今对超声图像分割的改进效果并不显著。针对上述情况，本文将SAM模型迁移到超声图像分割领域，利用PEFT技术，结合旁路CNN编码器和跨分支交互适配器来微调SAM，将SAM视为因果推断任务，构建因果模型来表示特征编码、提示和输出之间的因果关系，引入反事实提示来使模型专注于事实提示生成，提出一种结合反事实提示生

成与级联解码的SAM改进方法(SAM combined with counterfactual prompt generator and cascaded decoder, SAMCD)。本文所提出的SAMCD方法与SAM及其变体(MedSAM、SAMCT)的比较如图1所示。

本文的贡献包括:1)改进SAMCT的CNN编码器,使其作为旁路编码器,设计了一种新型简单的跨分支交互适配器(simple cross-branch interaction adapter, SCIA)模块,通过使用旁路CNN编码器分支以及SCIA模块,补充ViT(vision Transformer)编码器所缺乏的局部信息,以提高模型对细节的捕捉能力。2)构建因果模型来表示模型不同变量之间的因果关系,并引入因果学习的反事实干预(counterfactual intervention, CF)机制,使得模型专注于事实提示生成,加强提示生成器的学习能力,同时减少模型对高质量提示的依赖。3)为获得更丰富的边缘信息,设计了级联解码器,先利用SAM的原始解码器创建先验掩码,再使用加入边界注意力的Transformer解码器和像素解码器,使得模型能够理解丰富的边缘信息和优化分割结果。4)采用两阶段的训练策略以提高模型的分割性能并加快模型收敛,其中第1阶段为训练交互分割模型,第2阶段为训练加入提示生成器的自动分割模型。

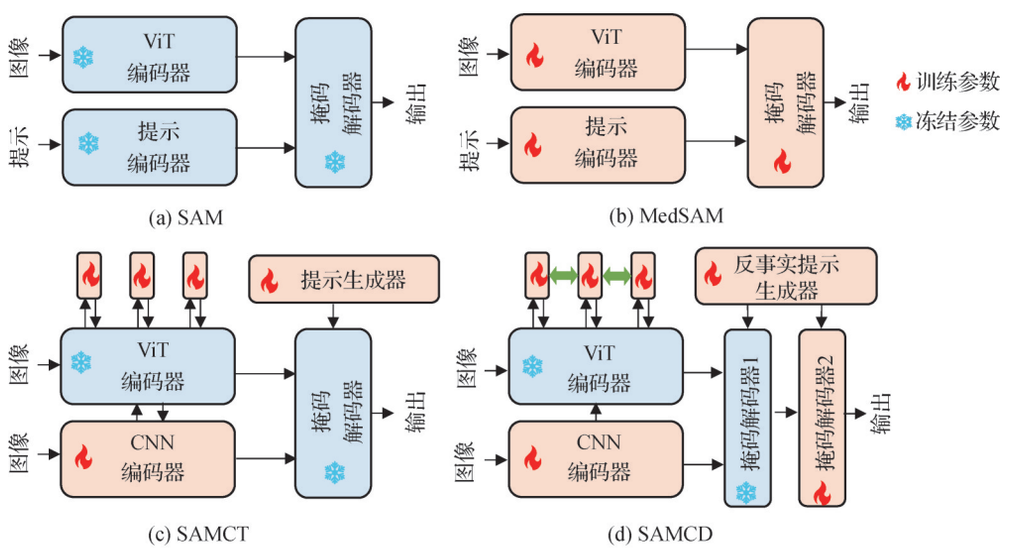


图1 SAMCD与SAM及变体的方法对比

Fig. 1 Comparison of SAMCD with SAM and related variants ((a) SAM; (b) MedSAM; (c) SAMCT; (d) SAMCD)

1 方 法

SAMCD整体结构如图2所示,主要由图像编码器、双解码器和提示生成器3部分组成,其中原始SAM的所有参数均被冻结。SAMCD图像编码器部分包括旁路CNN编码器与原始SAM的ViT编码器,两者并行运行且同时对输入图像进行编码,在编码器中引入SCIA模块进行微调,以促进旁路CNN图像编码器和ViT图像编码器之间的交互,使模型能捕获具有局部感受野的信息。解码器使用更复杂的级联解码器,首先利用SAM的原始解码器(即图2中解码器1)来创建先验掩码,用于指导后续更复杂的解码;然后在第2次解码中使用加入边界注意力的Transformer解码器和像素解码器,即图2中解码器2,用于优化分割结果。为了取代SAM所需的提示输入以实现自动分割,SAMCD引入提示生成器,通过将旁路CNN编码器提取的特征进行多尺度特征融合来生成提示并输入到解码器,最终输出期望的分割二值图像。在模型训练阶段利用反事实干预学习来获得效果更好的提示生成器,而在模型推理阶段通过提示生成器来生成提示,以实现全自动分割。

1.1 图像编码器

超声图像相比于自然图像,存在复杂的病变边界和多样的形态特征,为了将SAM模型迁移到超声图像分割领域,同时减少计算资源的消耗,本文将SAMCT原有ViT编码器中的适配器替换为参数共享的适配器(Wang等,2024),改进CNN编码器分支,使其作为旁路编码,设计了一个轻量化的跨分支交互适配器SCIA模块来补充ViT编码器所缺少的局

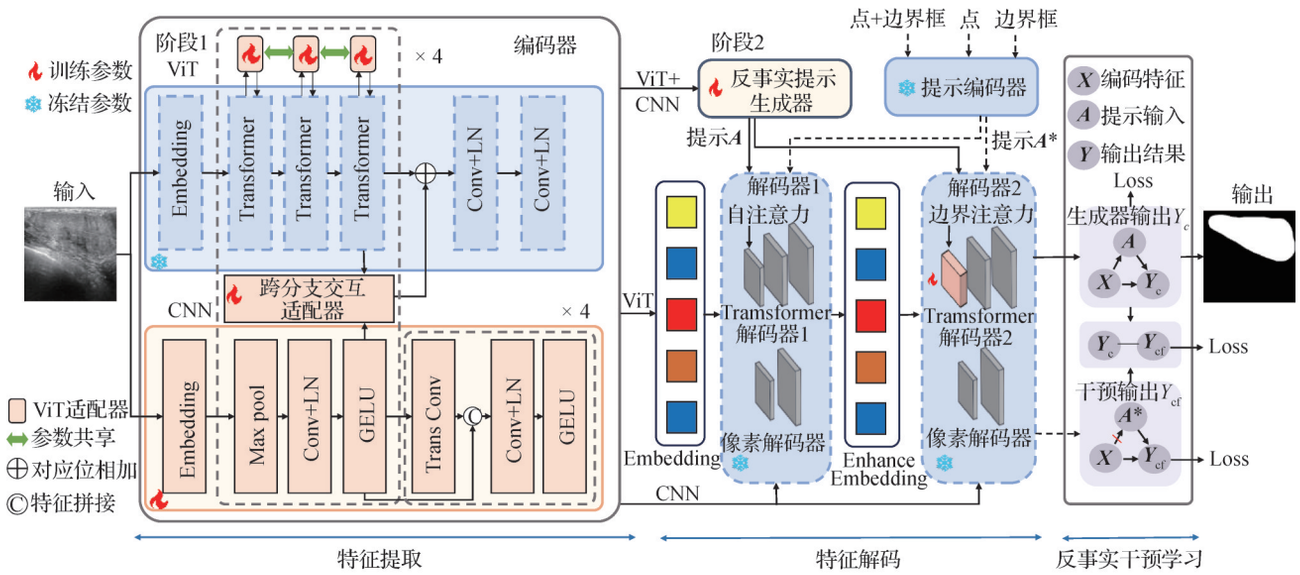


图2　SAMCD整体结构示意图

Fig. 2　Schematic diagram of the overall structure of SAMCD

部信息。

ViT编码器中参数共享的适配器如图3所示，主要由降维层和升维层组成，该设计允许两个相邻的适配器之间分享参数，以减少参数量，提高适配器的效率。

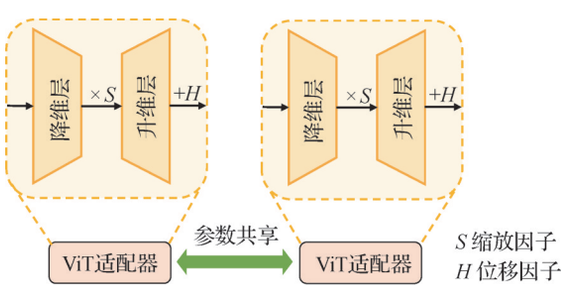


图3　ViT适配器模块结构示意图

Fig. 3　Schematic diagram of ViT adapter module structure

旁路CNN编码器如图2所示，包括特征提取块和特征扩展块，由最大池化层（max pooling layer，MP）、卷积层（convolutional layers，Conv）、层规范化（layer normalization，LN）、高斯误差激活函数（Gaussian error linear unit，GELU）和转置卷积（transposed convolution，TC）组成。对于CNN图像嵌入 $\boldsymbol{F}_{\text{cnn}} \in \mathbf{R}^{H\times W\times C}$，首先通过4个特征提取块，获得特征图 $\boldsymbol{F}_{16} \in \mathbf{R}^{\frac{H}{16}\times\frac{W}{16}\times C}$；之后通过特征扩展块，应用转置卷积，通过跳跃连接将转置卷积的输出与来自特征提取块相同分辨率特征图连接起来，经过卷积、层规范化和激活函数得出不同分辨率的特征图 $\boldsymbol{F}_{32} \in \mathbf{R}^{\frac{H}{8}\times\frac{W}{8}\times 8C}$、$\boldsymbol{F}_{64} \in \mathbf{R}^{\frac{H}{4}\times\frac{W}{4}\times 4C}$、$\boldsymbol{F}_{128} \in \mathbf{R}^{\frac{H}{2}\times\frac{W}{2}\times 2C}$ 和 $\boldsymbol{F}_{256} \in \mathbf{R}^{H\times W\times C}$。特征提取和特征扩展过程可以表述为

$$\boldsymbol{F}_{\text{c}} = \sigma(LN(Conv(MP(\boldsymbol{F}_{\text{cnn}})))) \tag{1}$$

$$\boldsymbol{F}_{\text{m}} = \sigma(LN(Conv(C_{\text{Trans}}(\boldsymbol{F}_{\text{c}}) \| \boldsymbol{F}_{\text{c}}))) \tag{2}$$

式中，MP表示最大池化操作；$Conv$表示卷积模块；LN表示层规范化；σ表示GELU激活函数，$\boldsymbol{F}_{\text{c}}$表示CNN特征提取的图像特征；$\|$表示沿通道维度的连接操作；$C_{\text{Trans}}$表示转置卷积模块；$\boldsymbol{F}_{\text{m}}$表示不同分辨率的特征图输出。

在图像编码器中使用跨分支交互适配器，SCIA模块结构如图4所示，其输入包括分别来自CNN分支与ViT分支的图像特征 $\boldsymbol{F}_{\text{c}}$ 和 $\boldsymbol{F}_{\text{v}}$。具体来说，首先将最大池化和平均池化应用于 $\boldsymbol{F}_{\text{v}}$，以生成粗略的空间注意力图 $\boldsymbol{F}_{\text{v}}'$；之后 $\boldsymbol{F}_{\text{c}}$ 通过交叉注意力建立与 $\boldsymbol{F}_{\text{v}}$ 依赖关系，$\boldsymbol{F}_{\text{v}}$ 作为查询（即 $\boldsymbol{Q}$），$\boldsymbol{F}_{\text{c}}$ 作为键（即 $\boldsymbol{K}$）和值（即 $\boldsymbol{V}$）；最终空间注意力图与融合后的特征进行线性组合并经过降维、激活与升维操作，以补充从CNN分支到ViT分支的局部信息。整个过程可以表述为

$$\boldsymbol{F}_{\text{v}}' = \sigma_2(Conv(MP(\boldsymbol{F}_{\text{v}}) \| AP(\boldsymbol{F}_{\text{v}}))) \tag{3}$$

$$\boldsymbol{F}_{\text{c}} = U(\sigma(D(CA(\boldsymbol{F}_{\text{v}}, \boldsymbol{F}_{\text{c}}, \boldsymbol{F}_{\text{c}})) + \boldsymbol{F}_{\text{v}}'))) \tag{4}$$

式中，AP表示平均池化操作；σ_2表示sigmoid激活函数；CA表示交叉注意力；D表示降维层；U表示升维层。

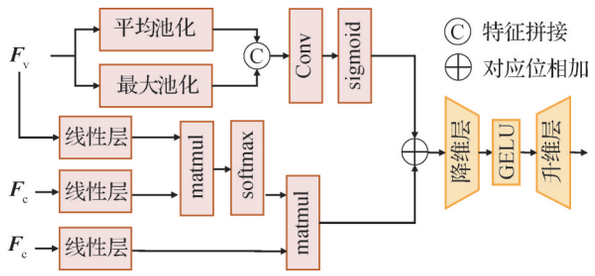


图4　SCIA模块结构示意图

Fig. 4　Schematic diagram of the structure of SCIA module

1.2　反事实干预下的提示生成器

在使用SAM时，用户需要向解码器输入点击或边界框来进行交互式图像分割。为了实现基于SAM的自动超声图像分割，本文所提出的SAMCD方法通过引入提示生成器来取代用户输出和提示编码器（Lin等，2024），提示生成器结构如图5所示。具体来说，提示生成器接受任务相关的正负指示符、CNN图像编码器的多分辨率图像嵌入，以及ViT图像编码器的图像嵌入$\boldsymbol{F}_{\text{vit}}$，图像嵌入经过线性层后，被转化为任务特定的特征空间，这些特征空间形成一系列新的图像嵌入$\boldsymbol{F}_{\text{enc}}$。基于这些任务相关的指标和图像嵌入，提示生成器可以生成正负点击和边界框的组合张量，以用于提示分割。

受Dou等人（2023）的启发，SAMCD在提示生成器的训练过程中引入反事实干预学习。如图2所示，通过构建因果模型来表示特征编码$\boldsymbol{X}$、提示输入$\boldsymbol{A}$和输出结果$\boldsymbol{Y}$之间的因果关系。其中因果链为$\boldsymbol{X}\rightarrow\boldsymbol{Y}$、$\boldsymbol{X}\rightarrow\boldsymbol{A}$和$\boldsymbol{X}\rightarrow\boldsymbol{Y}\leftarrow\boldsymbol{A}$。在理想情况下，提示$\boldsymbol{A}$应仅依赖于$\boldsymbol{X}$并与其共同预测$\boldsymbol{Y}$。然而，由于$\boldsymbol{X}$中存在各种混杂因素，从而混淆网络的学习过程。为此，本文引入反事实干预以切断混杂因素对模型的影响。

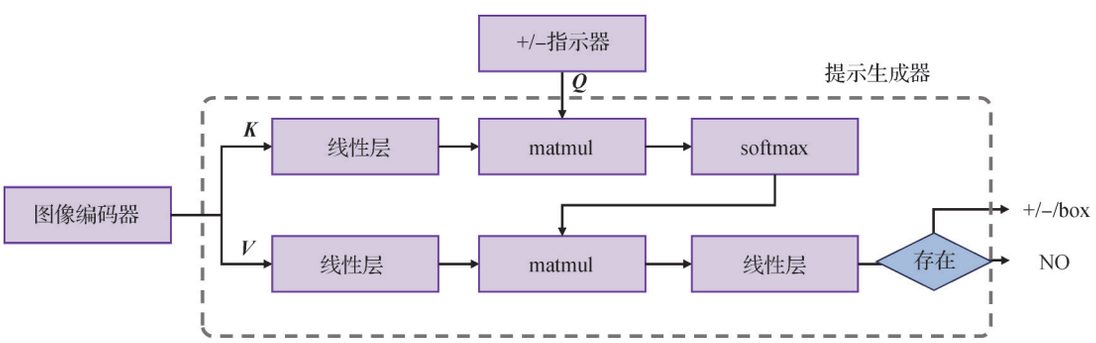


图5　提示生成器结构示意图

Fig. 5　Schematic diagram of the structure of prompt generator

反事实干预具体实现过程如下：

首先，通过提示生成器生成一组张量，产生相应的输出$\boldsymbol{Y}_{\text{f}}$；接着，引入提示干预$\boldsymbol{p}$，干预的提示是通过向图像背景区域输入随机位置的点来生成，并生成干预后的输出张量$\boldsymbol{Y}_{\text{cf}}$。将$\boldsymbol{Y}_{\text{cf}}$与原始输出$\boldsymbol{Y}_{\text{f}}$两者对应位相减后取绝对值，计算原始输出和干预输出之间的差异。通过计算$\boldsymbol{Y}_{\text{cf}}$、$\boldsymbol{Y}_{\text{f}}$和$\left|\boldsymbol{Y}_{\text{f}}-\boldsymbol{Y}_{\text{cf}}\right|$与GT的损失来进行一致性约束。整个过程定义为

$$\boldsymbol{T}_{\text{prompt}}=\boldsymbol{\Phi}_{\text{a-gen}}(\boldsymbol{F}_{\text{enc}}) \tag{5}$$

$$\boldsymbol{T}_{\text{rand}}=\boldsymbol{\Phi}_{\text{p-enc}}(\boldsymbol{p}) \tag{6}$$

$$\boldsymbol{Y}_{\text{f}}=\boldsymbol{\Phi}_{\text{dec2}}(\boldsymbol{\Phi}_{\text{dec1}}\left(\boldsymbol{F}_{\text{vit}},\boldsymbol{F}_{\text{cnn}},\boldsymbol{T}_{\text{prompt}}\right),\boldsymbol{F}_{\text{cnn}},\boldsymbol{T}_{\text{prompt}}) \tag{7}$$

$$\boldsymbol{Y}_{\text{cf}}=\boldsymbol{\Phi}_{\text{dec2}}(\boldsymbol{\Phi}_{\text{dec1}}\left(\boldsymbol{F}_{\text{vit}},\boldsymbol{F}_{\text{cnn}},\boldsymbol{T}_{\text{rand}}\right),\boldsymbol{F}_{\text{cnn}},\boldsymbol{T}_{\text{rand}}) \tag{8}$$

$$L=Loss(\left|\boldsymbol{Y}_{\text{f}}-\boldsymbol{Y}_{\text{cf}}\right|,\boldsymbol{Y}_{\text{f}},\boldsymbol{Y}_{\text{cf}};\boldsymbol{GT}) \tag{9}$$

式中，$\boldsymbol{F}_{\text{enc}}$表示图像编码器的图像嵌入；$\boldsymbol{T}_{\text{prompt}}$和$\boldsymbol{T}_{\text{rand}}$分别表示生成的提示嵌入和干预提示的嵌入；$\boldsymbol{F}_{\text{vit}}$和$\boldsymbol{F}_{\text{cnn}}$表示ViT和CNN编码器的图像嵌入；$\boldsymbol{Y}_{\text{f}}$和$\boldsymbol{Y}_{\text{cf}}$分别表示提示生成器、干预提示的输出结果；$L$表示模型的损失函数监督。值得注意的是，反事实干预学习仅在自动分割模型训练过程中使用，而在推理过程中没有引入反事实的干预提示，使用的是经过反事实训练得到的提示生成器来生成提示，以实现全自动分割。

1.3　双解码器

原始SAM的解码器是由Transformer解码器和像素解码器组成。Transformer解码器处理从图像编码器中提取的图像嵌入，并利用自注意力机制（self-attention，SA）评估各个图像区域的重要性，再通过交叉注意机制对相关区域进行分割。随后，利用像素解码器细化此输出，生成详细的分割图。作为Transformer解码器的补充，SAM的原始像素解码器直接将图像嵌入上采样为尺寸$(H/4)\times(W/4)$的分割

图。然而，由于该解码器分辨率较低，因此难以捕获超声医学图像分割中的复杂局部细节和小尺寸医疗目标。

受Cheng等人(2024)的启发，本文提出的SAMCD方法引入更复杂的级联解码来捕获局部细节，如图2所示。在第1次解码时，SAMCD使用SAM的原始解码器创建先验掩码，为第2次解码提供基础。第2次解码使用改进的Transformer解码器和原始像素解码器来优化第1次解码的分割结果。相较于原始的Transformer解码器，解码器2将自注意力机制改为边界注意力机制。边界注意力模块结构如图6所示，通过边界注意力机制和先验掩码的信息引导，捕获超声图像复杂的病变边界和多样的形态特征，从而增强分割过程。边界注意力机制处理过程可以表示为

$$\boldsymbol{X} = \boldsymbol{M}\left(f_{\text{softmax}}\left(\boldsymbol{K}\boldsymbol{Q}^{\text{T}}\right)\boldsymbol{V} + \boldsymbol{X}\right) \tag{10}$$

式中，$\boldsymbol{X}$是Transformer的输入查询特征；$\boldsymbol{K}$、$\boldsymbol{Q}$和$\boldsymbol{V}$是交叉注意力中的键、查询和值；$\boldsymbol{M}$是解码器1的输出结果。

旁路CNN编码器输出的特征依次被输入到解码器1和解码器2，从而增强整体解码过程。解码器2接收解码器1生成的先验掩码以及编码器的特征，以做进一步的细化，解码器2能够学习边界注意力中更多概率图的信息，并且可以为不同的前景区域分配不同程度的重要性，从而准确地捕捉图像中的细节和边界，生成更精确的分割结果。整个模型的解码过程可表示为

$$\boldsymbol{F}_{\text{en}} = \boldsymbol{\Phi}_{\text{dec1}}(\boldsymbol{F}_{\text{img}}, \boldsymbol{F}_{\text{cnn}}, \boldsymbol{T}_{\text{prompt}}) \tag{11}$$

$$\boldsymbol{M} = \boldsymbol{\Phi}_{\text{dec2}}(\boldsymbol{F}_{\text{en}}, \boldsymbol{F}_{\text{cnn}}, \boldsymbol{T}_{\text{prompt}}) \tag{12}$$

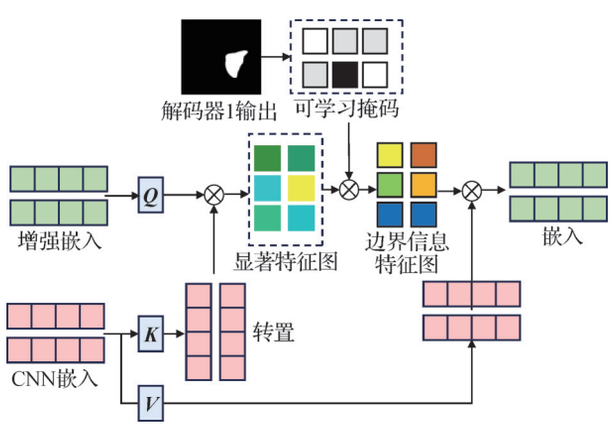


图6　边界注意力模块结构示意图

Fig. 6　Schematic diagram of the structure of the edge attention module

式中，$\boldsymbol{F}_{\text{img}}$表示中间图像特征；$\boldsymbol{T}_{\text{prompt}}$表示提示编码器编码的稀疏提示标记；$\boldsymbol{F}_{\text{en}}$表示解码器1输出；$\boldsymbol{M}$对应于预测的掩码。

1.4　训练策略

SAMCD采用两阶段的训练策略，以提高模型的分割性能，加快模型收敛。其中第1阶段训练交互分割模型，旁路CNN图像编码器与SAM中的ViT图像编码器并行工作，以补充局部特征；在级联解码时，由SAM的原始解码器生成先验概率掩码，指导第2次更复杂的解码过程，以提高模型捕捉细粒度局部细节的能力；第2阶段训练是在第1阶段的基础上，训练加入提示生成器的自动分割模型，通过引入反事实干预来加强其学习过程。

交互分割模型的提示选择策略为：当存在前景时，训练过程中的提示包含1个随机采样的正点、1个随机采样的负点和1个随机移动的边界框；否则，提示只包含2个随机取样的负点。两阶段训练均采用复合损失函数监督。复合损失函数表示为

$$L_{\text{stage1}} = \alpha L_{\text{Dice}} + \beta L_{\text{BCE}} \tag{13}$$

$$L_{\text{stage2}} = \alpha L_{\text{Dice}} + \beta L_{\text{BCE}} + L_{\text{SmoothL1}} \tag{14}$$

式中，L_{Dice}表示Dice损失；L_{BCE}表示交叉熵损失；L_{SmoothL1}表示平滑L_1损失函数；α和β为平衡权重。

2　实　验

2.1　数据集与评价指标

为评估所提出的SAMCD方法在超声医学图像分割任务上的效果和通用性，本文选择TN3K(thyroid nodule 3K)(Gong等，2023)和BUSI(breast ultrasound image)(Al-Dhabyani等，2020)这两个公开的标准数据集。其中，TN3K由3 493幅图像组成，并带有像素级甲状腺结节注释，该数据集是在南方医科大学朱江医院收集。BUSI乳腺超声数据集由780幅图像组成，并带有像素级乳腺癌注释，包括良性、恶性和正常图像，由于正常图像中不存在待分割的目标，故本文去除BUSI中的133幅正常图像。TN3K的划分参考Lin等人(2023)的工作；BUSI随机分成7∶1∶2，分别用于训练、验证和测试。表1给出了实验中使用的两个数据集的统计信息。

本文选择医学图像分割常用的指标进行评价，包括Acc(accuracy)、HD(Hausdorff distance)、DSC

表1　数据集统计信息

Table 1　Statistic information of the datasets

/幅

数据集	图像数量	训练集	测试集	验证集
TN3K	3 493	2 303	614	576
BUSI	647	454	129	64

(Dice similarity coefficient)、mIoU(Mean intersection over union)、敏感性(sensitivity, Sen)和特异性(specificity, Spe)。除HD外,其余指标的取值范围均为[0,1],且越接近1时模型分割效果越好。HD没有固定取值范围,其值越小时分割效果越好。为比较不同方法的时间复杂度以及设备资源开销,选择参数量(parameters)和复杂度(GFLOPs)作为指标。

2.2　实验设置

实验硬件平台为Intel(R)Xeon(R)Silver 4214R CPU @ 2.40 GHz, NVIDIA GeForce RTX 3090, 编程语言为Python3.9, 深度学习框架为PyTorch。网络训练时设置Batch size为4, 学习率为0.000 1,训练轮数为200,选用Adam优化器。在训练之前,SAMCD使用在SA-1B上训练的权重初始化,即从SAM继承网络参数,其余参数采用随机初始化。在训练过程中,所有图像都调整为256×256像素分辨率,即将宽度和高度小于256的图像填充为零值的边缘,使用双线性插值来调整图像的大小。

2.3　实验结果与分析

2.3.1　损失函数对比实验

在训练阶段均使用由Dice损失和交叉熵损失构成的复合损失函数来计算网络预测与实际的差距。首先尝试4组复合损失函数权重组合,以确定合适的权值,在BUSI数据集上的实验结果如表2所示,训练中损失值变化曲线如图7所示。

表2　不同损失函数权重组合下的验证指标对比

Table 2　Comparison of validation metrics under different combinations of loss function weights

α	β	DSC/%	mIoU/%	HD/mm	Acc/%
1	0	84.11	75.73	26.71	96.84
0.8	0.2	**84.29**	**76.18**	**25.13**	97.06
0.5	0.5	82.90	73.94	27.38	**97.12**
0.2	0.8	79.86	70.04	32.45	96.38

注:加粗字体表示各列最优结果。

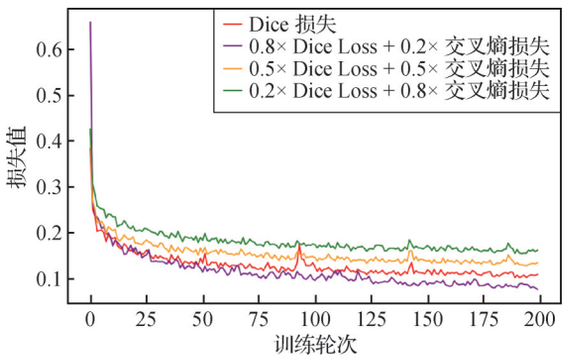


图7　第1阶段复合损失函数损失值变化曲线

Fig. 7　Loss value variation curves of composite loss function in the first stage

由表2可以得出,相对于单纯使用Dice损失,加入小权值的交叉熵损失(即β = 0.2),使得整体模型分割Dice值提高0.18%,当β权值增大后,模型开始出现性能下降。由图7可知,当选用0.8倍Dice损失搭配0.2倍交叉熵损失的复合函数(即α = 0.8且β = 0.2)时,初始损失较低并且收敛速度快于其他组合,因此后续实验选用该权重组合。

2.3.2　分割效果比较

选择SAM和相关变体的方法(MedSAM、SAMed和SAMCT),以及当前主流的基于CNN的方法和基于Transformer的方法(U-Net、CE-Net、SwinUnet、TransFuse和TransUNet),与本文的SAMCD方法进行对比,展示所提方法的优越性。表3—表4中的数据为各模型的平均值。

表3为在TN3K数据集上与其他分割方法的实验结果对比。可以看出,在TN3K数据集上本文SAMCD在DSC、mIoU、HD、Acc、Sen、Spe评价指标上取得最优或次优的评估结果。部分模型以现有公开性能数据作为参照(Lin等,2023),与SAM以及相关变体的方法相比,SAMCD的DSC值达到83.66%,比原始SAM提高45.22%,比MedSAM、SAMed、SAMCT的DSC值分别提高8.04%、2.93%和0.73%。HD值达到29.01 mm,比SAMed和SAMCT分别降低1.46 mm、0.74 mm。在mIoU和Acc上相较于SAMCT分别提升0.91%(74.89% vs 73.98%)和0.16%(97.23% vs 97.07%),相较于其余3种对比模型,平均提升20.43%和12.91%。因此,SAMCD相较于SAM相关方法有着更好的性能和更小的训练参数量,在计算量低于SAMCT的情况下,实现更优的分割效果。

与基于CNN的方法和基于Transformer的方法

相比，SAMCD的DSC值分别比U-Net、CE-Net、SwinUnet、TransFuse、TransUNet高4.65%、3.29%、13.58%、5.16%和2.22%；在Acc上比TransFuse和TransUNet高0.79%、0.29%；在Sen和Spe上平均高出5种方法4.95%、0.46%。本文相较于SAM相关方法有着更好的性能和更小的参数量。SAMCD在计算量仅略高于U-Net、SwinUnet和TransFuse的情况下，其分割效果显著优于这些分割模型，相较于TransUNet这一类复杂模型，SAMCD在保持较少训练参数的同时，仍能展现出优异的分割性能。

表3　不同网络模型在TN3K数据集上的结果对比

Table 3　Comparison of results of different network models on TN3K dataset

网络模型	训练参数量/M	GFLOPs	DSC/%	mIoU/%	HD/mm	Acc/%	Sen/%	Spe/%
U-Net(Ronneberger等,2015)	31	54.60	79.01	69.40	34.14	96.44	84.80	97.91
CE-Net(Gu等,2019)	29	35.78	80.37	70.66	32.79	96.40	85.91	97.44
SwinUnet(Cao等,2023)	27	5.91	70.08	58.19	44.13	94.94	79.34	96.55
TransFuse(Zhang等,2021)	26	11.53	78.50	68.60	30.98	96.44	82.06	97.95
TransUNet(Chen等,2021)	105	50.70	81.44	72.05	30.98	96.94	87.13	97.75
SAM(Kirillov等,2023)	0	743.98	38.44	27.34	83.26	60.55	75.16	59.86
MedSAM(Ma等,2024)	636	743.98	75.62	64.71	53.26	95.75	83.32	96.75
SAMed(Zhang和Liu,2023)	108	260.80	80.73	71.31	30.47	96.64	86.34	97.58
SAMCT(Lin等,2024)	62	73.71	82.93	73.98	29.75	97.07	86.45	**98.27**
SAMCD(本文)	32	71.32	**83.66**	**74.89**	**29.01**	**97.23**	**88.76**	97.98

注：加粗、下划线字体表示各列最优、次优结果。

表4为在BUSI数据集上与其他分割方法的实验结果对比。可以看出，在BUSI数据集上本文SAMCD方法在DSC、mIoU、HD、Sen评价指标上均优于其他分割方法。SAMCD的DSC值达到84.29%，与SAM以及相关变体的方法相比，比原始SAM提高30.3%，比MedSAM、SAMed、SAMCT的Dice值分别提高4.56%、7.23%和0.9%；SAMCD的HD值达到25.13 mm，比SAMed和SAMCT降低9.14 mm、3.71 mm，在mIoU和Acc上相较于SAMCT，SAMCD分别提升1.56%(76.18% vs 74.62%)和0.35%(97.06% vs 96.71%)，相较于其余3种对比模型，平均提升17.54%和6.07%。与基于CNN的方法和基于Transformer的方法相比，SAMCD的DSC值分别比U-Net、CE-Net、SwinUnet、TransFuse、TransUNet高出6.18%、2.69%、17.06%、10.77%和2.07%；在Acc上比TransUNet略差，比TransFuse高出0.85%；在Sen上平均高出5种方法6.79%。

综合来看，由于超声图像分割任务区域目标具有不确定性，而U-Net、SwinUnet侧重于CNN捕获信息，模型特性导致其不适用于超声图像这类随机性强的分割任务。MedSAM虽然在大规模医学数据集上训练，但其超声图像数据相对较少，且没有边缘信息的补充。TransUNet和TransFuse侧重于结合Transformer和CNN之间的优势，可以更好地捕获图像的特征信息。CE-Net和SAMCT是针对皮肤病分割和CT图像分割任务提出的，其图像都存在不确定的边界区域问题，迁移到超声分割任务中展现出较好的模型泛化性。本文SAMCD是在SAMCT基础上的改进，同时SAMCD模型更加轻量化，在不使用任何外部提示的情况下具有更高的分割性能，取得当前先进水平的结果，具有很强的竞争力。

2.3.3　适配器消融实验

为验证ViT编码器中适配器对模型参数量和性能的影响，针对该适配器模块进行消融实验。首先，在保持SAMCD其他模块不变的基础上，将其原始的适配器更换为参数共享的适配器(Share)。适配器消融实验结果如表5所示。可以看出，模型使用SAMCT中的适配器相较于不使用适配器，在TN3K和BUSI数据集上的DSC分别提升0.75%和0.98%，但参数量和GFLOPs也有所增加。在使用参数共享适配器和在无适配器时的参数量的差异可以忽略不计，且在TN3K和BUSI数据集上DSC分

表4　不同网络模型在BUSI数据集上的结果对比

Table 4　Comparison of results of different network models on BUSI dataset

网络模型	训练参数量/M	GFLOPs	DSC/%	mIoU/%	HD/mm	Acc/%	Sen/%	Spe/%
U-Net(Ronneberger等,2015)	31	54.60	78.11	69.60	33.60	96.69	83.16	97.95
CE-Net(Gu等,2019)	29	35.78	81.60	73.62	29.19	96.86	81.72	98.09
SwinUnet(Cao等,2023)	27	5.91	67.23	55.58	47.02	95.28	78.94	96.75
TransFuse(Zhang等,2021)	26	11.53	73.52	64.28	34.95	96.21	78.61	97.69
TransUNet(Chen等,2021)	105	50.70	82.22	<u>74.77</u>	<u>27.54</u>	**97.26**	83.19	**98.69**
SAM(Kirillov等,2023)	0	743.98	53.99	41.55	130.72	81.25	64.51	81.83
MedSAM(Ma等,2024)	636	743.98	79.73	70.28	33.54	95.83	86.57	97.42
SAMed(Zhang和Liu,2023)	108	260.80	77.06	66.13	34.27	95.89	85.57	97.43
SAMCT(Lin等,2024)	62	73.71	<u>83.39</u>	74.62	28.84	96.71	<u>87.31</u>	97.93
SAMCD(本文)	32	71.32	**84.29**	**76.18**	**25.13**	<u>97.06</u>	**87.92**	<u>97.96</u>

注:加粗、下划线字体表示各列最优、次优结果。

表5　ViT编码器中适配器模块的消融实验

Table 5　Ablation experiments of the adapter module in the ViT encoder

适配器	参数量/M	GFLOPs	TN3K数据集			BUSI数据集		
			DSC/%	mIoU/%	HD/mm	DSC/%	mIoU/%	HD/mm
无	32.62	71.32	82.16	72.70	29.49	82.90	73.94	27.38
SAMCT(Lin等,2024)	36.17	72.23	82.91	73.75	29.02	83.88	75.29	26.71
Share(Wang等,2024)	32.63	71.32	**83.66**	**74.89**	**29.01**	**84.29**	**76.18**	**25.13**

注:加粗字体表示各列最优结果。

别提升1.5%和1.39%,优于使用SAMCT适配器的分割结果。

因此,使用SAMCT适配器在模型性能上虽然有所提升,但增加了计算量。相比之下,参数共享适配器在保持参数量基本不变的情况下,能够实现更高的性能提升。

2.3.4　不同模块消融实验

为验证SAMCD中提出的各个模块对网络整体性能的影响与贡献,通过消融实验系统地分析各模块在超声医学图像分割任务上的效果。消融实验包括跨分支交互适配器模块、解码器2和反事实干预的不同组合,这些模块被依次引入到带有参数共享适配器和旁路CNN编码器的SAM网络架构中,并在TN3K和BUSI数据集上进行训练和评估。消融实验结果如表6所示,可以看出在实验2和实验3中,即使只引入SAMCD的单个模块(SCIA或解码器2),也会带来明显的性能提升,在TN3K数据集上的DSC相比基线模型分别提升2.37%(SCIA)和2.92%(解码器2);在BUSI数据集上分别提升2.25%(SCIA)和5.33%(解码器2),这表明跨分支交互SCIA模块和解码器2是有效的。实验4同时添加SCIA和解码器2,可实现性能的进一步提升。这一阶段的优化验证了级联解码策略在处理复杂医学图像中的有效性,尤其是在精确识别边缘和小区域方面。实验5则同时引入反事实干预,这一策略旨在通过模拟潜在的错误预测来加强模型对真实情况的适应性和鲁棒性。相比于实验4(未引入反事实干预),实验5在TN3K和BUSI数据集上的DSC提升0.5%和0.18%,从而验证了反事实干预在提升模型泛化能力和纠正错误预测方面的潜力。

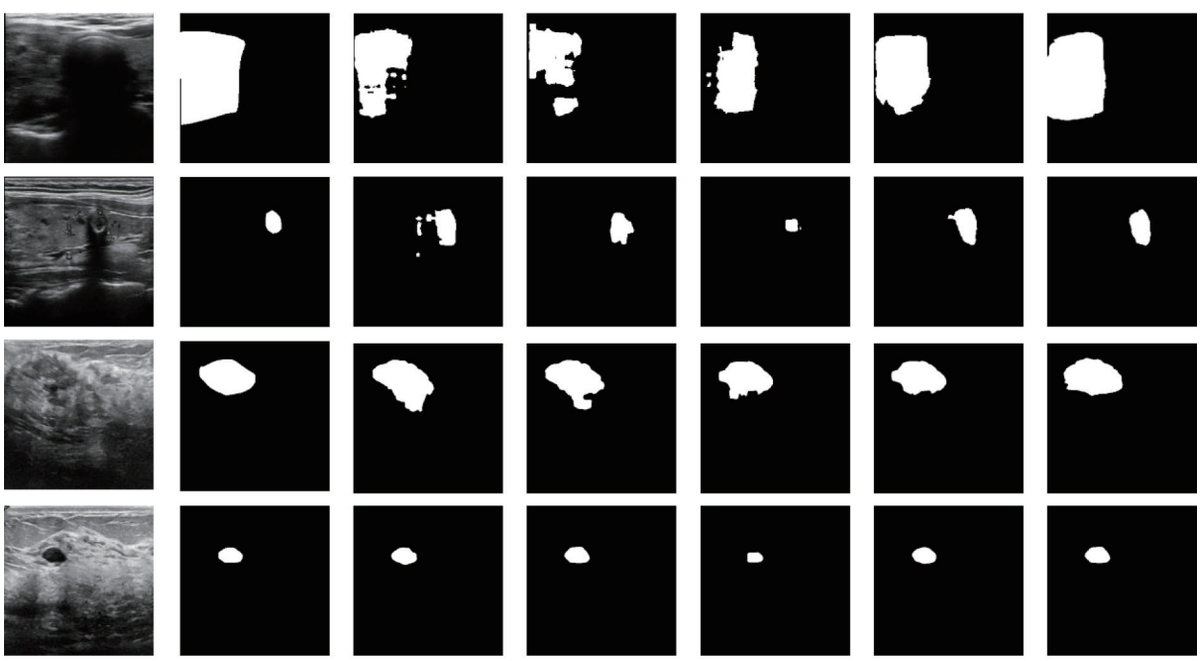
综上,通过组合跨分支交互适配器、解码器2和反事实干预模块,SAMCD均达到最佳的分割性能,充分验证了这一复合模型在超声医学图像分割任务上的优越性。图8展示了不同模块消融实验的可视化

表6　SAMCD不同模块的消融实验结果

Table 6　Ablation study on different component combinations of SAMCD

实验	SCIA	解码器2	CF	TN3K数据集			BUSI数据集		
				DSC/%	mIoU/%	HD/mm	DSC/%	mIoU/%	HD/mm
1	–	–	–	78.34	68.31	32.78	76.09	63.33	33.62
2	√	–	–	80.71	70.96	31.03	78.34	68.91	32.45
3	–	√	–	81.26	71.72	30.50	81.42	71.99	28.49
4	√	√	–	83.16	73.70	29.49	84.11	75.56	26.75
5	√	√	√	**83.66**	**74.89**	**29.01**	**84.29**	**76.18**	**25.13**

注：加粗、下划线字体表示各列最优、次优结果，“√”表示采用，“–”表示未采用。



(a) 原图　(b) 图像标签　(c) 实验1　(d) 实验2　(e) 实验3　(f) 实验4　(g) 实验5

图8　不同模块的消融实验可视化结果

Fig. 8　Ablation visualization results ((a) original images; (b) ground truth; (c) experiment 1; (d) experiment 2; (e) experiment 3; (f) experiment 4; (g) experiment 5)

结果，当同时使用SCIA和解码器2时，模型可以更好地捕捉目标的范围与边界细节，表现出对目标前景的精确定位能力；在引入反事实干预后，模型在目标分割的准确性和细节处理上得到了进一步的提升。

2.3.5　提示生成器消融实验

为验证提示生成器的有效性，在TN3K和BUSI数据集上通过不同方式的提示对SAMCD模型的性能进行测试评估，涉及模拟人工提示的单次点击(1p)、单个边界框(Bbox)、点击加边界框(1p + Bbox)，以及非人工提示下由提示生成器所生成的提示，包括采用反事实与未采用反事实两种方式。实验结果如表7所示，可以看出使用单次点击时，DSC值在TN3K数据集上达到79.62%，在BUSI数据集上为77.28%。这种单点交互虽然简单，但在进行基础分割任务时展示出一定的有效性；当引入单个边界框作为提示时，模型的性能有显著提升，DSC值在TN3K和BUSI上分别达到85.89%和86.32%。这说明边界框作为提示能更准确地界定目标区域，从而提高分割的准确性；结合单次点击和边界框的方法进一步提升了模型的性能，使得DSC值在TN3K和BUSI上分别提升到89.06%和91.13%，表明结合使用多种类型的提示可以显著增强模型的分割效果；

表7　提示生成器的消融实验结果

Table 7　Prompt generator ablation experiments

提示方式	TN3K数据集				BUSI数据集			
	DSC/%	mIoU/%	HD/mm	Acc/%	DSC/%	mIoU/%	HD/mm	Acc/%
1p	79.62	69.87	32.48	96.47	77.28	66.48	33.62	95.69
Bbox	85.89	77.57	**24.04**	96.55	86.32	76.76	23.68	97.01
1p + Bbox	**89.06**	**81.18**	24.76	**98.14**	**91.13**	**84.19**	**23.12**	**98.41**
提示生成器(无反事实)	83.16	73.70	29.49	97.01	84.11	75.56	26.75	96.84
反事实提示生成器	83.66	74.89	29.01	97.23	84.29	76.18	25.13	97.06

注:加粗字体表示各列最优结果。

通过使用反事实提示生成器,SAMCD的DSC值在TN3K和BUSI数据集上分别达到83.66%和84.29%,mIoU分别达到74.89%和76.18%,较未使用反事实的提示生成器均有所提高。这表明提示生成器能够有效地模拟人工交互操作,其生成的提示可达到与少量人工交互相当的分割性能水平。

为直观展示是否采用反事实干预的提示生成器所生成的提示张量间差异,随机选择50幅测试样本进行对比。由于人工提示下的点击加边界框提示能够生成最优结果,因此以该种人工提示方式为基准,分别计算由提示生成器所生成的两种提示张量与人工提示张量的余弦相似度。该值越高表明越接近于点击加边界框这种人工提示,实验结果如图9所示。可以看出,添加反事实干预训练的提示生成器生成的提示更接近人工提示的点击加边界框提示,故其能够生成精度更高的提示输出。

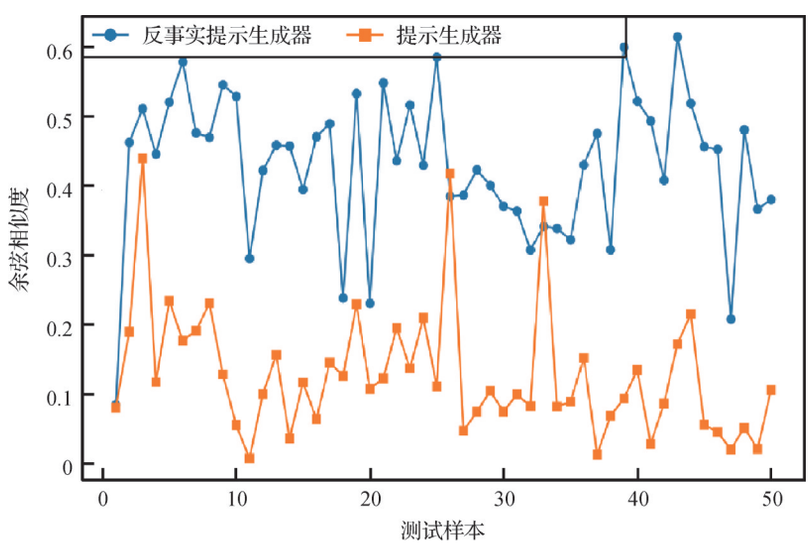


图9　是否采用反事实生成训练的提示生成张量相似度对比

Fig. 9　Comparison of prompt generator tensor similarity with and without counterfactual training

2.3.6　不同方法的可视化分析

为了更直观地展示SAMCD方法的有效性,在TN3K和BUSI数据集上对不同方法的可视化结果进行展示。图10给出7种方法的可视化结果比较,包括传统的U-Net、CE-Net、TransUNet方法,以及基于SAM的方法,如SAM、MedSAM、SAMCT与本文SAMCD。其中,第1—2行展示BUSI数据集的结果,第3—5行则是TN3K数据集的结果。由图10可以看出,U-Net和CE-Net方法虽然能够大体识别出感兴趣的区域,但在细节上如边缘的精确划分和小区域的捕获上存在不足;TransUNet虽然相比于U-Net在结构理解上有所提升,但在捕捉极小的或相互交织的病变时,仍会漏检或误判;SAM对于边缘模糊或大小不一的多病变结构,性能不很稳定,会忽略掉

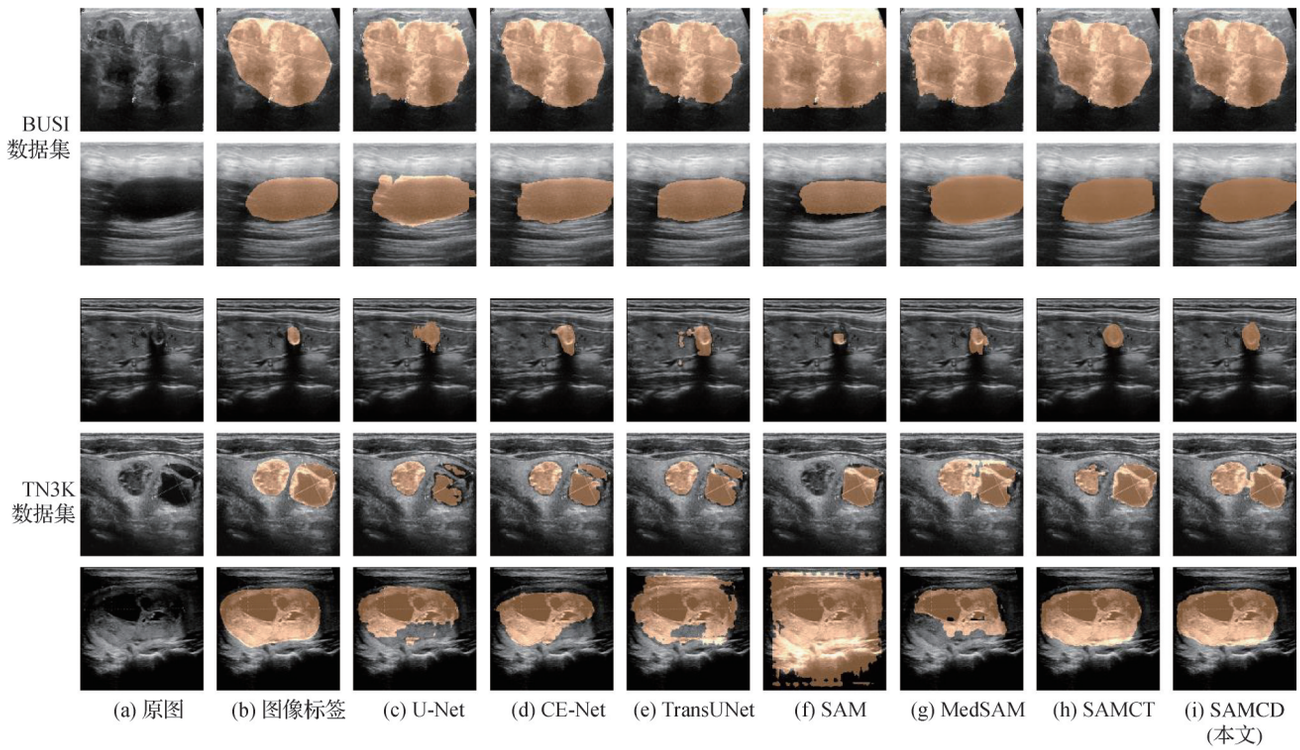


图 10　不同方法的可视化结果对比

Fig. 10　Visualization of comparative experiments ((a) original images; (b) ground truth; (c) U-Net; (d) CE-Net; (e) TransUNet; (f) SAM; (g) MedSAM; (h) SAMCT; (i) SAMCD(ours))

一些重要的细节信息；MedSAM 在处理边缘细节和小病变区域方面表现较为突出；SAMCT 的上下文感知能力加强模型对整体图像结构的理解，从而在维持区域一致性和分割精度方面超越基础的 SAM 和 MedSAM；SAMCD 方法在处理复杂图像时具有优势，尤其在 TN3K 数据集上，该方法不仅准确捕获目标区域的形状和位置，还细致地描绘出边缘和病变区域的界限。

综上，与现有方法相比，SAMCD 提供的分割结果更为精细和接近真实标记，显示出与真实分割的高度一致性，确保超声医学图像分割在实际应用中的准确性和可靠性，从而更好地支持临床决策和疾病诊断。

3　结　论

超声医学图像具有低对比度、边界模糊和形状复杂的特点，这导致直接利用 SAM 进行分割易丢失边缘信息，造成分割准确率不高，且需要人工干预进行半自动的交互式分割。针对以上问题，本文在 SAMCT 基础上进行改进，提出 SAMCD 方法。首先，改进 CNN 分支，设计一种轻量化的跨分支交互适配器 SCIA 模块和串行的解码器，以解决边缘信息不足的问题；其次，通过级联解码来捕获超声图像分割中的复杂局部细节和小目标，获得多尺度增强边缘特征，以达到增强边缘信息的目的；此外，通过提示生成器实现全自动分割，并引入反事实干预来加强提示生成器的学习。在 TN3K 和 BUSI 数据集上进行训练和测试，SAMCD 的 DSC 值分别达到 83.66% 和 84.29%，并与现有的 9 种先进方法进行测试比较，实验结果表明所提 SAMCD 明显优于其他方法，具有更高的分割性能，从而验证 SAMCD 在超声医学图像分割任务上的有效性。SAMCD 的网络参数虽然比 SAMCT 有所减少，计算速度也有所提高，但是对于目前的临床应用设备而言，计算所需的内存仍然较大，还需进一步降低模型的参数量。目前 SAMCD 旨在解决从背景中提取前景的二值分割问题，未来可以扩展到解决复杂和多样的医学图像多类分割任务。

参考文献(References)

Al-Dhabyani W, Gomaa M, Khaled H and Fahmy A. 2020. Dataset of

breast ultrasound images. Data in Brief, 28: #104863 [DOI: 10.1016/j.dib.2019.104863]

Cao H, Wang Y Y, Chen J, Jiang D S, Zhang X P, Tian Q and Wang M N. 2023. Swin-Unet: Unet-like pure Transformer for medical image segmentation//Proceedings of 2023 European Conference on Computer Vision. Tel Aviv, Israel: Springer: 205-218 [DOI: 10.1007/978-3-031-25066-8_9]

Chen J Y, You H J and Li K. 2020. A review of thyroid gland segmentation and thyroid nodule segmentation methods for medical ultrasound images. Computer methods and programs in biomedicine, 185: #105329 [DOI:10.1016/j.cmpb.2020.105329]

Chen J N, Lu Y Y, Yu Q H, Luo X D, Adeli E, Wang Y, Lu L, Yuille A L and Zhou Y Y. 2021. TransUNet: Transformers make strong encoders for medical image segmentation [EB/OL]. [2024-08-20]. https://arxiv.org/pdf/2102.04306.pdf

Chen L X, Lin C C, Zheng Z L, Mo Z F, Huang X Y and Zhao G S. 2023. Review of Transformer in computer vision. Computer Science, 50(12): 130-147 (陈洛轩, 林成创, 郑招良, 莫泽枫, 黄心怡, 赵淦森. 2023. Transformer在计算机视觉场景下的研究综述. 计算机科学, 50(12): 130-147) [DOI: 10.11896/jsjkx.221100076]

Cheng J L, Ye J, Deng Z Y, Chen J P, Li T B, Wang H Y, Su Y Z, Huang Z Y, Chen J L, Jiang L, Sun H, He J J, Zhang S T, Zhu M and Qian Y. 2023. SAM-Med2D [EB/OL]. [2024-08-20]. https://arxiv.org/pdf/2308.16184.pdf

Cheng Z H, Wei Q Y, Zhu H R, Wang Y, Qu L Q, Shao W and Zhou Y Y. 2024. Unleashing the potential of SAM for medical adaptation via hierarchical decoding//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 3511-3522 [DOI: 10.1109/CVPR52733.2024.00337]

Dhruv P and Naskar S. 2020. Image classification using convolutional neural network (CNN) and recurrent neural network (RNN): a review//Proceedings of ICMLIP 2019 on Machine Learning and Information Processing. Singapore, Singapore: Springer: 367-381 [DOI: 10.1007/978-981-15-1884-3_34]

Dou H Z, Zhang P Y, Su W, Yu Y L, Lin Y N and Li X. 2023. GaitGCI: generative counterfactual intervention for gait recognition//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE, 2023: 5578-5588 [DOI: 10.1109/CVPR52729.2023.00540]

Fu Z H, Yang H R, So A M C, Lam W, Bing L D and Collier N. 2023. On the effectiveness of parameter-efficient fine-tuning//Proceedings of the 37th AAAI Conference on Artificial Intelligence. Washington, USA: AAAI Press: 12799-12807 [DOI: 10.1609/aaai.v37i11.26505]

Gong H F, Chen J X, Chen G Q, Li H F, Li G B and Chen F. 2023. Thyroid region prior guided attention for ultrasound segmentation of thyroid nodules. Computers in Biology and Medicine, 155: #106389 [DOI: 10.1016/J.COMPBIOMED.2022.106389]

Gu Z W, Cheng J, Fu H Z, Zhou K, Hao H Y, Zhao Y T, Zhang T Y, Gao S H and Liu J. 2019. CE-Net: context encoder network for 2d medical image segmentation. IEEE Transactions on Medical Imaging, 38(10): 2281-2292 [DOI: 10.1109/tmi.2019.2903562]

Han K, Wang Y H, Chen H T, Chen X H, Guo J Y, Liu Z H, Tang Y H, Xiao A, Xu C J, Xu Y X, Yang Z H, Zhang Y M and Tao D C. 2023. A survey on vision Transformer. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(1): 87-110 [DOI: 10.1109/tpami.2022.3152247]

Kirillov A, Mintun E, Ravi N, Mao H Z, Rolland C, Gustafson L, Xiao T T, Whitehead S, Berg A C, Lo W Y, Dollár P and Girshick R. 2023. Segment anything//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE: 3992-4003 [DOI: 10.1109/ICCV51070.2023.00371]

Lin X, Xiang Y Y, Wang Z H, Cheng K T, Yan Z Q and Yu L. 2024. SAMCT: segment any CT allowing labor-free task-indicator prompts [EB/OL]. [2024-08-20]. https://arxiv.org/pdf/2403.13258.pdf

Lin X, Xiang Y Y, Zhang L, Yang X, Yan Z Q and Yu L. 2023. SAMUS: adapting segment anything model for clinically-friendly and generalizable ultrasound image segmentation [EB/OL]. [2024-08-20]. https://arxiv.org/pdf/2309.06824v1.pdf

Liu Z L, Chen G, Shan Z Y and Jiang X Q. 2018. Segmentation of spine CT image based on deep learning. Computer Applications and Software, 35(10): 200-204, 273 (刘忠利, 陈光, 单志勇, 蒋学芹. 2018. 基于深度学习的脊柱CT图像分割. 计算机应用与软件, 35(10): 200-204), 273) [DOI: 10.3969/j.issn.1000-386x.2018.10.036]

Ma J, He Y T, Li F F, Han L, You C Y and Wang B. 2024. Segment anything in medical images. Nature Communications, 15(1): #654 [DOI: 10.1038/s41467-024-44824-z]

Noble J A and Boukerroui D. 2006. Ultrasound image segmentation: a survey. IEEE Transactions on Medical Imaging, 25(8): 987-1010 [DOI: 10.1109/tmi.2006.877092]

Ronneberger O, Fischer P and Brox T. 2015. U-Net: convolutional networks for biomedical image segmentation//Proceedings of the 18th International Conference on Medical Image Computing and Computer-Assisted Intervention. Munich, Germany: Springer: 234-241 [DOI: 10.1007/978-3-319-24574-4_28]

Shi P L, Qiu J N, Abaxi S M D, Wei H, Lo F P W and Yuan W. 2023. Generalist vision foundation models for medical imaging: a case study of segment anything model on zero-shot medical segmentation [EB/OL]. [2024-08-20]. https://arxiv.org/pdf/2304.12637.pdf

Wang M, Huang Z Z, He H G, Lu H C, Shan H M and Zhang J P. 2024. Potential and prospects of segment anything model: a survey. Journal of Image and Graphics, 29(6): 1479-1509 (王淼, 黄智忠, 何晖光, 卢湖川, 单洪明, 张军平. 2024. 分割一切模型SAM的潜力与展望: 综述. 中国图象图形学报, 29(6): 1479-1509) [DOI: 10.11834/jig.230792]

Wang Y, Ge X K, Ma H, Qi S L, Zhang G J and Yao Y D. 2021. Deep learning in medical ultrasound image analysis: a review. IEEE Access, 9: 54310-54324 [DOI: 10.1109/access.2021.3071301]

Wang Y N, Chen K, Yuan W M, Tang Z P, Meng C and Bai X Z. 2024. SAMIHS: adaptation of segment anything model for intracranial hemorrhage segmentation//Proceedings of 2024 IEEE International Symposium on Biomedical Imaging. Athens, Greece: IEEE: 1-5 [DOI: 10.1109/ISBI56570.2024.10635673]

Wu J D, Ji W, Liu Y P, Fu H Z, Xu M, Xu Y W and Jin Y M. 2023. Medical SAM adapter: adapting segment anything model for medical image segmentation [EB/OL]. [2024-08-20]. https://arxiv.org/pdf/2304.12620.pdf

Xia F, Shao H J and Deng X. 2022. Cross-stage deep-learning-based MRI fused images of human brain tumor segmentation. Journal of Image and Graphics, 27(3): 873-884 (夏峰, 邵海见, 邓星. 2022. 融合跨阶段深度学习的脑肿瘤MRI图像分割. 中国图象图形学报, 27(3): 873-884) [DOI: 10.11834/jig.210330]

Xiao H G, Li L, Liu Q Y, Zhu X H and Zhang Q H. 2023. Transformers in medical image segmentation: a review. Biomedical Signal Processing and Control, 84: #104791 [DOI: 10.1016/J.BSPC.2023.104791]

Yu D, Peng Y J and Guo Y F. 2023. Ultrasonic image segmentation of thyroid nodules-relevant multi-scale feature based h-shape network. Journal of Image and Graphics, 28(7): 2195-2207 (于典, 彭延军, 郭燕飞. 2023. 面向甲状腺结节超声图像分割的多尺度特征融合"h"形网络. 中国图象图形学报, 28(7): 2195-2207) [DOI: 10.11834/jig.220078]

Zhang K D and Liu D. 2023. Customized segment anything model for medical image segmentation [EB/OL]. [2024-08-20]. https://arxiv.org/pdf/2304.13785.pdf

Zhang Y D, Liu H Y and Hu Q. 2021. TransFuse: fusing Transformers and CNNs for medical image segmentation//Proceedings of the 24th International Conference on Medical Image Computing and Computer Assisted Intervention. Strasbourg, France: Springer: 14-24 [DOI: 10.1007/978-3-030-87193-2_2]

作者简介

霍一儒，男，硕士研究生，主要研究方向为医学图像分割。
E-mail: 1202210005@student.stdu.edu.cn
封筠，通信作者，女，教授，主要研究方向为计算机视觉、机器学习和复杂网络分析。E-mail: fengjun@stdu.edu.cn
刘娜，女，硕士研究生，主要研究方向为医学图像分割。
E-mail: 1202310006@student.stdu.edu.cn
史屹琛，男，博士研究生，主要研究方向为深度学习。
E-mail: shiyichen@sjtu.edu.cn
殷梦莹，女，硕士研究生，主要研究方向为计算机视觉。
E-mail: 2438185425@qq.com