利用基于DINOv3的自监督特征实现鲁棒的少样本医学图像分割

Guoping Xu[1], Jayaram K. Udupa[2], Weiguo Lu[1], You Zhang[1#]

[1]医学人工智能与自动化实验室（MAIA），德克萨斯大学西南医学中心放射肿瘤学系，美国德克萨斯州达拉斯，邮编 75390[2]医学图像处理组（MIPG），宾夕法尼亚大学放射学系，美国宾夕法尼亚州费城，邮编 19104

[#]电子邮箱：You.Zhang@UTSouthwestern.edu

摘要

基于深度学习的自动医学图像分割在临床诊断和治疗规划中发挥着关键作用，但由于标注训练数据稀缺，在小样本场景下仍面临挑战。近年来，在大型自然图像数据集上训练的自监督基础模型（如 DINOv3）展现出强大的密集特征提取潜力，有望应对少样本学习的挑战。然而，领域差异阻碍了这些模型直接应用于医学图像。在本研究中，我们提出了一种新型框架 DINO-AugSeg，利用 DINOv3 特征应对少样本医学图像分割这一挑战。具体而言，我们引入了 WT-Aug，这是一种基于小波的特征级增强模块，通过扰动频率分量来丰富 DINOv3 提取特征的多样性；还引入了 CG-Fuse，这是一种上下文信息引导融合模块，利用交叉注意力将富含语义的低分辨率特征与空间细节丰富的高分辨率特征相结合。在涵盖五种成像模态（包括磁共振成像、计算机断层扫描、超声、内窥镜检查和皮肤镜检查）的六个公开基准上进行的大量实验表明，在样本有限的条件下，DINO-AugSeg 的表现始终优于现有方法。结果凸显了结合小波域增强与上下文融合对于实现鲁棒特征表示的有效性，表明 DINO-AugSeg 是推进少样本医学图像分割的一条有前景的方向。代码和数据将在以下网址公开：https://github.com/apple1986/DINO-AugSeg。

关键词：少样本学习、语义分割、小波、交叉注意力

1. 引言

医学图像分割旨在将图像中的每个像素（或体素）归类到特定的解剖结构或病灶类别。它是医学图像分析中的一项基础任务，在诊断 [1], 疾病监测 [2], 治疗规划 [3], 手术引导 [4], 和病理定量评估 [5]中发挥着关键作用。深度学习的兴起极大地推动了这一领域的发展，其中 U-Net 已成为最具影响力的架构之一。U-Net 在传统的编码器–解码器范式基础上引入跳跃连接，使多尺度上下文特征与细粒度特征得以融合，从而实现更精准的分割 [6]。其设计简洁高效，且在多种成像模态上表现出色，因此已成为现代医学图像分割研究的基石 [7]。在 U-Net 的基础上，研究者提出了一系列变体以进一步提升性能 [8]。例如，SegResNet 采用残差块来改善梯度流和特征表示 [9]；UNet++ 利用嵌套式密集跳跃连接，实现更有效的多尺度特征融合 [10]；nnU-Net 则提出了一种自适应、自配置框架，能够针对给定数据集自动优化预处理、网络架构、训练计划和后处理策略 [11]。总体而言，这些进展彰显了基于卷积的 U-Net 衍生架构不断演进的趋势，以及它们在医学图像分割中的核心作用。除了纯卷积架构之外，近期研究越来越多地将 Transformer 引入分割框架，原因在于 Transformer 能够捕捉图像区域之间的长程依赖，从而克服基于卷积的方法固有的局部性限制。TransUNet [12] 和 LeViT-UNet [13] 等混合方法将卷积骨干网络与 Transformer 模块相结合，以兼顾局部特征提取和全局上下文建模。相比之下，UNETR [14] 和 SwinUNETR [15], 等完全基于 Transformer 的架构利用自注意力机制，实现端到端体积图像分割，并增强上下文表示。

尽管这些特定任务方法在医学图像分割方面取得了显著进展，但仍有若干局限阻碍其应用于常规临床实践。其主要缺点在于依赖大型标注数据集进行监督训练。构建此类标注数据集既成本高昂又耗费人力，因为需要临床专家进行细致的逐层标注。受大语言模型（LLM）成功的启发，[16, 17], 自监督学习（SSL）通过利用无标签数据，已成为一种颇具前景的替代方案。这一范式转变推动了从小规模特定任务模型向大规模可泛化框架的转变，因为收集大量无标签数据通常比制作专家标注更为可行。

沿着这一发展方向，利用 SSL 在大规模自然图像上预训练的视觉基础模型，作为目标检测和语义分割等下游任务的特征提取器，展现出卓越的通用性和鲁棒性。其中，DINO 系列是一个具有代表性的研究方向 [18-20]。尤其是 DINOv3 [20], 采用新颖的 Gram 锚定策略，在 17 亿张图像上训练；与早期的自监督和弱监督基础模型（如 DINOv2 [19] 和掩码自编码器（MAE）[21],）相比，它提取高质量稠密特征表示的能力更强，因此在医学图像分割领域尤具潜力。例如，MedDINOv3 [22]研究了将 DINOv3 适配至医学成像领域的方法：采用与 DINOv3 相同的训练方案，在 SSL 机制下使用 387 万张轴位 CT 切片进行预训练。在四个分割基准上微调后，MedDINOv3 的表现达到或超越了最先进的方法。此外，近期研究 [23] 进一步表明，DINOv3 可作为强有力的基线，用于提取适用于多种医学成像模态的鲁棒先验特征。如图 1 所示，面对多种噪声扰动，DINOv3 展现出较强的特征鲁棒性。从 DINOv3 编码器生成的主成分分析（PCA）特征图在不同扰动类型下保持一致。值得注意的是，DINOv3 提取的特征（图 1 第三行）保留了分割对象清晰的结构轮廓，凸显了其在严苛条件下的可靠性。

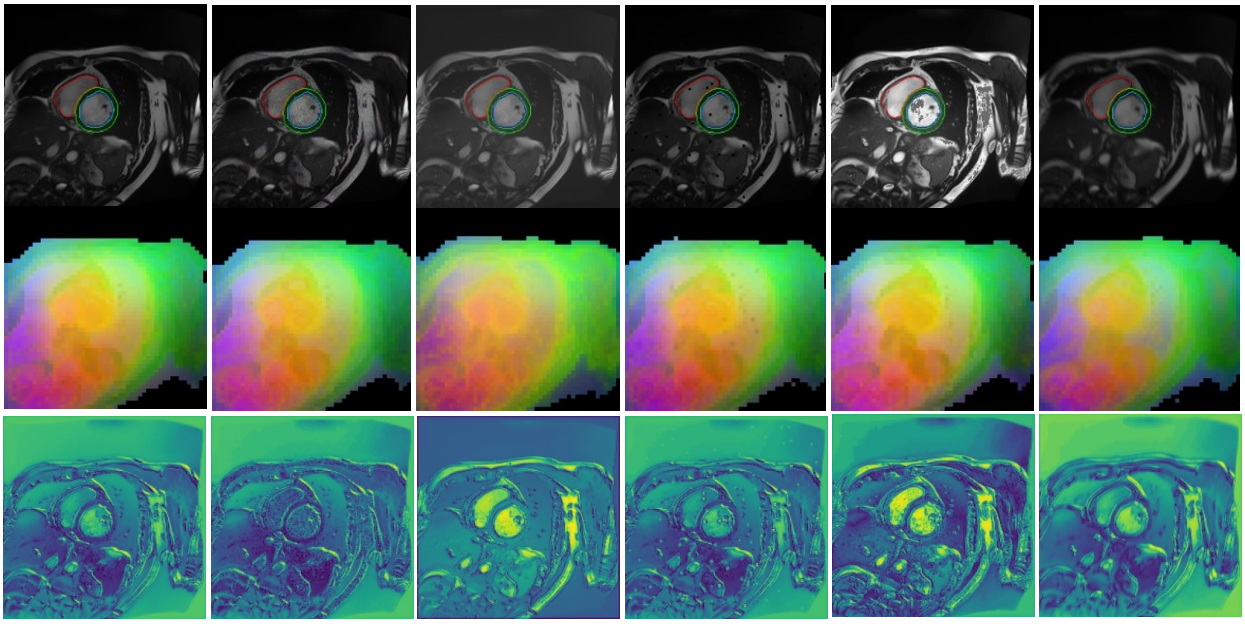


图 1. 第一行展示了心脏电影 MRI 数据集中的一个代表性切片（原始图像），以及经过不同扰动增强处理后的变体。边界以红色、绿色和蓝色叠加显示，分别对应右心室腔、心肌和左心室。第二行展示了从 DINOv3 图像编码器提取的主成分分析（PCA）特征图。第三行展示了第一行图像的特征表示，这些表示通过最大投影获得

在 DINOv3 的第一阶段沿通道维度进行。这些结果表明，DINOv3 提取的特征具有很强的表征能力，并且在面对各种破坏时依然稳健。

鉴于 DINOv3 即使在各种破坏下也能提取稳健且具有代表性的特征，我们可以预期，其稳定的特征提取性能将有助于各种下游图像分析任务。事实上，在无监督目标检测[24] 和多模态图像配准[25]中，使用 DINO 特征的有效性已得到验证。然而，当我们将 DINOv3 作为冻结编码器，与类 U-Net 解码器结合，用于 ACDC数据集 [26], 的医学图像分割时，如图2所示，随着训练样本数量从 20 减少到 2，验证集和测试集上的平均Dice分数都急剧下降。这引出了一个问题：即使使用 ***DINOv3*** 作为稳健的特征提取器，在训练数据有限时，为什么分割性能仍会显著下降？鉴于 DINOv3 编码器已经经过充分预训练，我们推测瓶颈可能在于随机初始化的解码器，它通常需要各种训练样本，才能稳健地学习并适应特定数据集。

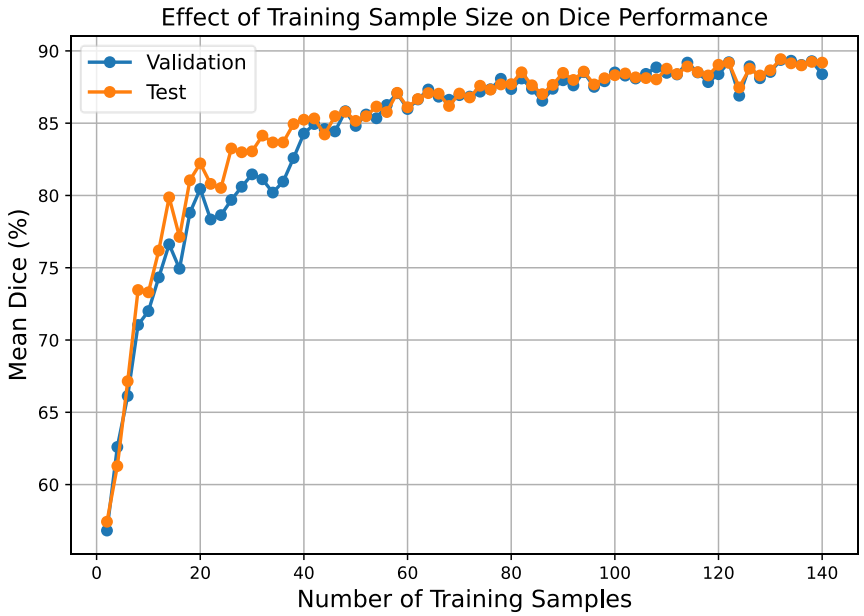


图2。使用 DINOv3 编码器（冻结）和类U-Net解码器时，ACDC数据集验证集和测试集上的Dice相似度分数随训练样本数量（2–140）的变化情况。当训练样本数量极为有限（从20个减少到2个）时，性能会大幅下降；随着训练样本数量增加，分割性能则稳步提升。

根据我们的假设，解码器需要多样化的训练样本才能有效地泛化。在小样本场景中，一种直接的策略是应用图像级增强 [27]。然而，传统的图像级增强（如噪声注入、随机网格失真和亮度

调整）可能收效有限，因为 DINOv3 提取的特征在很大程度上对这类扰动保持不变（如图 1 所示）。在这种情况下，DINOv3 起到了“噪声滤波器”的作用，不利地降低了增强数据的多样性，进而削弱了训练所得解码器的泛化能力。这一观察促使本研究探讨：在标注数据有限的情况下，相较于图像级增强，特征级增强能否更有效地提升基础模型辅助的少样本医学图像分割性能。图像级增强在 DINOv3 的“噪声过滤”之前进行，而特征级增强则在 DIVOv3 完成特征提取之后进行，因此不受过滤效应影响。然而，直接对 DINO 特征应用经典空间域增强可能会损害特征完整性：例如，随机运动可能扭曲对象结构（图3，第一行），随机网格可能破坏纹理一致性（图3，第一行和第三行），而泊松噪声或随机亮度则可能掩盖关键信息或降低特征可辨别性（图3，第二行）。

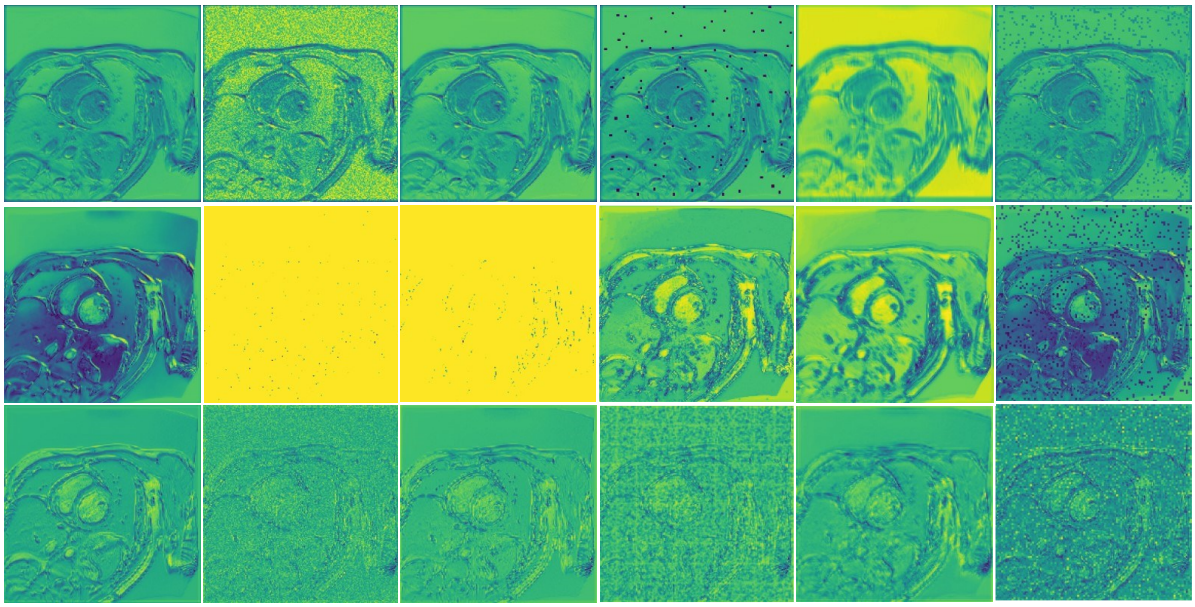


图3。对 DINOv3 编码器第一阶段生成的特征进行增强。各行依次对应第一个通道（顶部）、跨通道最大池化（中部）和跨通道平均池化（底部）。

从另一角度来看，特征级空间域增强可能会显著改变数据集的特征分布。如图4所示，我们采用 t-SNE [28] 和 UMAP [29] 可视化了不同策略下的增强特征，并观察到与原始未增强特征分布（图4中以星号标记）相比，存在明显的分布偏移。这引出了第二个关键问题：如何设计特征级增强以避免出现剧烈的分布偏移

同时缓解医学图像分割中小样本训练所面临的挑战？ 为此，我们研究了小波域中的特征级增强，该方法能够对频率分量进行独立操控。这一特性实现了良好的权衡：在引入有意义变化的同时，不会严重扭曲特征表示，从而在标注数据有限的情况下提升少样本分割效果。图3（最后一列，小波域增强）和图4（粉色簇）提供了支持证据。具体而言，小波域增强能够保留结构和纹理信息（图3），同时使特征分布与原始样本保持接近（图4）。与直接扰动整幅图像或特征图的空间域增强不同，小波域增强会对各个小波分量进行独立、随机的操作。因此，增强后的特征能够更好地保留分割对象的底层结构信息。

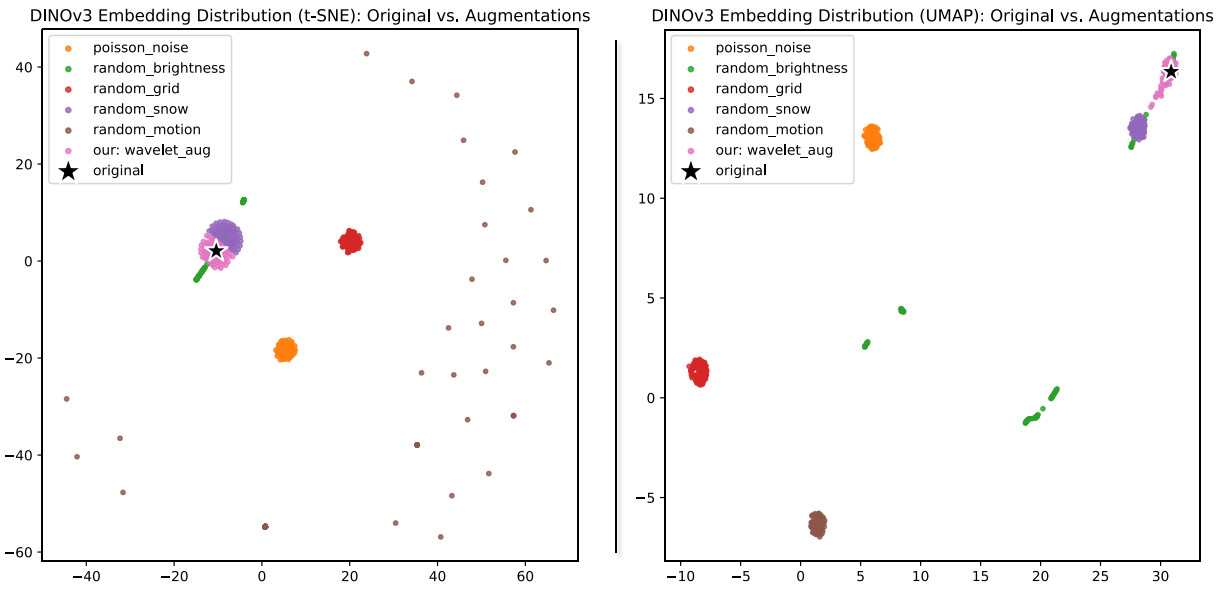


图4。 使用 t-SNE（左）和 UMAP（右）可视化 DINOv3 编码器输出的特征分布（全局一维池化嵌入）。为简化起见，可视化时直接对最终输出嵌入应用增强。实际上，分割任务中的特征级增强通常应用于 DINOv3 不同阶段的二维特征图。

除特征级增强之外，另一个同等关键的挑战在于设计有效的特征融合策略，为解码器提供语义丰富且具有代表性的信息，从而提升分割性能。U-Net 和 nnU-Net [11] 通过直接拼接融合编码器和解码器特征，将对象特定表征的学习留给后续解码器层在训练过程中完成。尽管这一策略简单且有效，但对于融合 DINO 衍生特征而言可能并非最优，因为后者已经包含了高层上下文语义 [19,

30]。一种更具原则性的方法是利用这些上下文线索来引导融合过程。为此，我们受原始 Transformer [31]的启发，提出了一种基于交叉注意力机制的 上下文引导特征融合模块（*CG-Fuse*）。在我们的设计中，解码器特征 [1]被构建为查询，而编码器特征则被转换为键和值。其根本动机是利用嵌入 DINO 特征中的丰富上下文信息来引导特征融合，从而使解码器能够更有效地关注对准确分割至关重要的对象级信息。

总而言之，本研究在利用基于DINOv3的自监督特征实现稳健的少样本医学图像分割方面提出了两项关键创新：（***1***）采用有效的小波域特征级增强，缓解标注训练样本稀缺以及DINOv3 编码器的“噪声过滤”效应所带来的限制；（***2***）利用DINOv3捕获的丰富高层上下文语义来引导特征融合，并提升解码器的分割性能。基于这两项进展，我们的主要贡献如下：

（a）我们提出了一种用于DINOv3特征的新型基于小波的特征级增强方法（*WT-Aug*）。（b）我们设计了一个上下文引导特征融合模块（*CG-Fuse*），以利用DINOv3特征中的高层上下文信息。

（c）我们提出了一种新型分割框架*DINO-AugSeg*，在编码器–解码器架构中集成了*WT-Aug*和*CG-Fuse*。

（d）我们在涵盖五种医学成像模态的六个数据集上进行了全面实验，包括磁共振、计算机断层扫描、超声、内窥镜检查和皮肤镜检查。结果表明，所提出的方法行之有效，尤其适用于少样本分割场景。

2. 相关工作

（1）图像级和特征级增强

数据增强通过从有限的原始数据中生成经过修改的样本，人为扩充训练数据集，已被证明是缓解深度学习中数据稀缺问题的有效策略 [32]。除了亮度调整等传统图像级技术以及

[1] 在此，我们将 DINOv3 最终层的特征指定为查询，因为它们包含最高层级的语义信息，并用于引导早期 DINOv3 层特征的融合。

随机旋转外，针对深度学习设计的高级增强方法也展现出显著成效 [33, 34]。这些方法不仅能对模型进行正则化，还可以通过鼓励模型重建缺失内容，拓展至自监督学习，从而促使网络学习更具判别性的特征。基于这一思路，掩码自编码器（MAE）[21] 会随机遮蔽图像块，并训练模型重建缺失内容。MAE 已证明，随机掩码可以作为一种可扩展且有效的大规模视觉模型预训练范式。受其成功启发，研究者提出了许多用于自监督学习的基于掩码的数据增强策略 [35-37]。此外，这类基于掩码的数据增强策略也已应用于医学图像分割领域 [38-40], ，并展现出促进代表性特征学习的显著效果。

除图像级增强外，特征级增强也被探索为缓解训练数据稀缺局限性的一种有效策略。在 [41], 中，有研究引入了一种随机子采样方法，用于在编码器模块内增强特征。在 [42], 中，通过应用随机卷积权重，并采用固定加权方案将所得特征与原始特征融合，实现了特征增强，从而促进了域泛化医学图像分割。UniMatch [43] 和 UniMatch V2 [44] 采用了逐通道Dropout策略 [45] ，在半监督分割框架中执行特征级增强，引入扰动以提高对无标签数据的利用率。尽管这些方法具有较好的效果，但它们局限于空间域，可能会无意中破坏原始特征的内在分布（见图4）。为解决这一局限性，我们研究了小波域中的特征级增强，并将其应用于少样本分割，利用小波域将特征分解为相互独立的频率分量的能力，从而引入受控变化，同时避免严重改变底层表征。

（2）用于图像分割的解码器

医学图像分割网络中的解码器旨在逐步对特征图进行上采样，使其恢复到原始分辨率，从而为每个目标结构生成最终的分割掩码 [8]。为了提高预测质量，充分利用编码器多个阶段的特征至关重要。U-Net、nnU-Net、U-Net++ [10]、DeepLabv3+ [46]和 PSPNet [47]等基于卷积的架构通过跳跃连接融合编码器特征与解码器特征。U-Net 和 nnU-Net 主要拼接来自对应编码器-解码器层、空间尺度相同的特征，而 U-Net++ 则通过密集连接的跳跃路径扩展了这一范式，可聚合来自

多个语义层级的特征。DeepLabv3+ 采用空洞卷积和多尺度上采样来捕获上下文信息，而 PSPNet 则整合金字塔池化，以组合不同尺度的表征。TransUNet [12] 和 LeViT-UNet[13], 等混合卷积-Transformer方法，利用卷积层进行局部特征提取、利用 Transformer 建模长程依赖的互补优势，通常在解码器中直接拼接这两类特征。此外，SegFormer [48] 和 SegDINO[49], 等纯基于Transformer的模型依靠融合来自不同编码器阶段的多层级特征，生成最终的语义预测。

与这些先前方法不同，后者主要通过沿跳跃连接直接拼接特征，并在解码器中进行融合；我们提出了一种上下文信息引导融合模块。该设计充分利用DINOv3更优越的语义表示，使解码器能够更有效地利用高层语义信息，从而提升分割性能。

(3) DINO系列与DINOv3在医学图像分割中的应用

DINO（无标签自蒸馏）是一种基于视觉Transformer（ViT）训练的自监督学习框架 [18]。一项关键发现是，与有监督ViT或卷积网络提取的特征相比，DINO提取的特征包含更丰富的语义信息，适用于密集预测任务 [50]。这些特征在目标分割等下游应用中展现出更强的迁移能力。在此基础上，DINOv2将该框架扩展至大规模基础模型训练，依托大量精选数据集，学习更稳健且泛化能力更强的视觉表示 [19]。最近推出的DINOv3则在数据准备和优化中采用了先进的自监督策略。值得注意的是，该模型提出了一种新颖的目标函数——Gram锚定，以减轻长时间训练过程中密集特征质量的下降 [20]。实证结果表明，DINOv3能够生成高质量且具有丰富语义信息的密集特征，在多种视觉任务中的表现优于自监督和弱监督基础模型。

由于具备强大的密集特征提取能力，基于 DINO 的架构已广泛应用于医学图像分割 [49, 51, 52]。例如，SegDINO 将冻结的DINOv3用作编码器骨干网络，并结合轻量级的基于MLP的解码器，在多种医学成像模态上均实现了稳定的最先进性能。同样，Dino U-Net [53], 采用编码器-解码器范式进行设计，利用 DINOv3 提供的高保真密集特征来提升分割准确率。总的来说，这些方法表明，语义丰富的密集

DINOv3 学习到的特征为医学图像分割提供了强大且通用的基础。

在本研究中，我们探究了 DINOv3 特征在少样本医学图像分割中的鲁棒性。与以往方法不同，我们的研究重点有两个方面：(1) 为 DINOv3 表征开发特征增强策略，以缓解训练样本有限带来的挑战；(2) 有效利用 DINOv3 特征中的上下文信息，引导解码器中的特征融合，从而提升最终的分割性能。

3. 方法

3.1 DINO-AugSeg 的整体架构

所提出的 DINO-AugSeg框架整体架构如图 5 所示。该框架由三个主要部分组成：（1）作为骨干网络的冻结的DINOv3 编码器；（2）通过 WT-Aug 模块增强的跳跃连接；（3）集成 CG-Fuse模块的解码器。

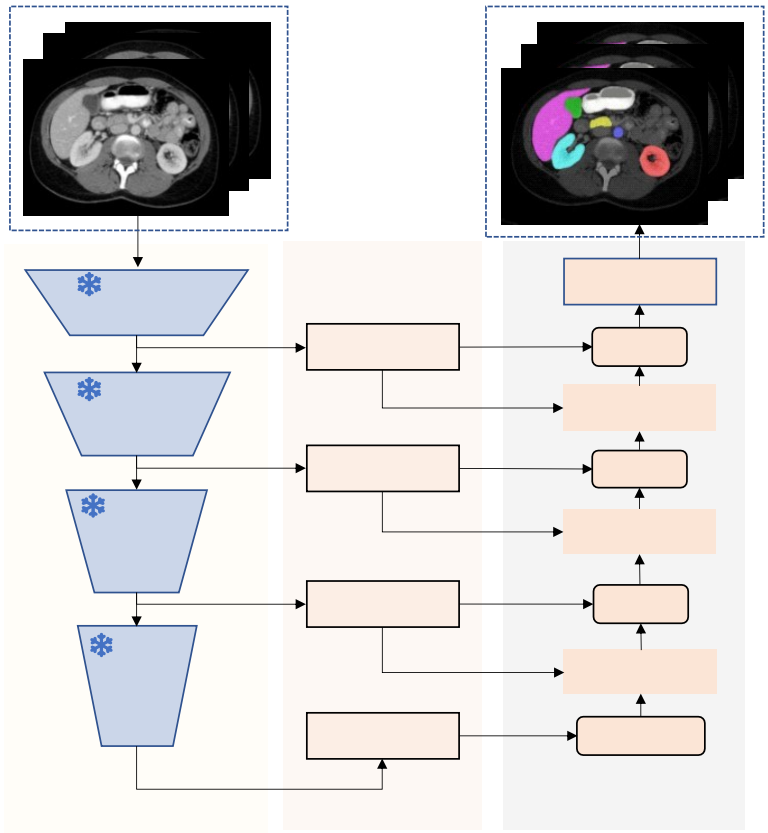


图 5. 所提出的 DINO-AugSeg框架的整体架构。该框架由一个用于多尺度特征提取的冻结的 DINOv3 编码器，以及应用于跳跃

连接，以及嵌入上下文引导融合（CG-Fuse）模块的解码器。CCU 表示通过转置卷积进行的拼接、卷积和上采样操作。

首先，将输入的医学图像送入冻结的DINOv3 编码器，并从其四个层级阶段中提取多尺度特征图。随后，通过相应的WT-Aug模块处理这些中间特征，在小波域中进行特征增强，以丰富表征的多样性。

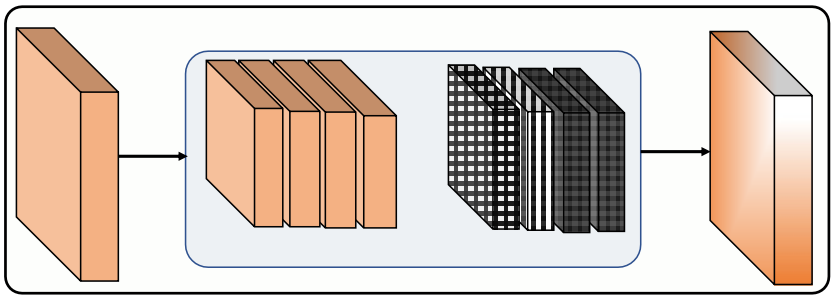
接下来，将增强特征传递至解码器。在此阶段，利用深层的低分辨率但上下文丰富特征，通过提出的CG-Fuse模块引导高分辨率特征的融合，从而有效整合语义信息与空间信息。然后，将CG-Fuse的输出特征与编码器同一阶段对应的特征一并送入CCU模块，该模块执行**拼**接、**卷**积以及基于转置卷积的**上**采样。

最后，将融合后的特征输入轻量级分割头（SegHead，如图5所示），生成最终的分割输出。

这种架构设计旨在利用DINOv3稳健的语义特征，同时引入特征增强和引导式融合策略，以提升分割准确率和泛化能力，尤其是在有限数据条件下。

3.2 基于小波的特征级增强

如图6所示，所提出的 WT-Aug 在小波域中进行特征级增强。它包含三个主要步骤：Haar 小波[2]分解[54], 频率分量增强，以及通过逆 Haar 小波变换进行重建。



[2] 本研究采用 Haar 小波变换，因为与其他小波族（例如 Daubechies、Symlets 等）相比，它更加简单且计算效率更高。

图6。所提出的 WT-Aug 模块的详细结构。首先，DINOv3 提取的多尺度特征（以棕色显示）通过 Haar 小波变换分解为四个频带：LL（低频近似）、LH（水平高频细节）、HL（垂直高频细节）和 HH（对角高频细节）。然后，将每个频带与四个随机生成的掩码逐元素相乘，以进行特征级增强。最后，通过逆 Haar 小波变换，将修改后的频带重建到空间特征域。

具体而言，首先使用 Haar 小波变换将 DINOv3 输出中某一阶段的特征图分解为四个独立的频率子带：LL（低频近似）、LH（水平高频细节）、HL（垂直高频细节）和 HH（对角高频细节）。这一过程可表示如下：

$$(LL, LH, HL, HH) = WT(F), \tag{1}$$

where *F* 表示输入特征图，*WT(.)* 表示 Haar 小波变换。

接下来，将每个频率子带与空间维度相同的随机生成掩码逐元素相乘，以生成增强后的频率分量：

$$(LL', LH', HL', HH') = (LL \odot M_{LL}, LH \odot M_{LH}, HL \odot M_{HL}, HH \odot M_{HH}), \tag{2}$$

其中，$\odot$ 表示逐像素相乘，M_* 表示随机掩码。

最后，通过逆 Haar 小波变换将增强后的频率分量重建到空间域：

$$F' = WT^{-1}(LL', LH', HL', HH'), \tag{3}$$

这种增强策略通过有选择地改变部分频率信息，同时保留目标的结构和纹理完整性，有效引入随机扰动。因此，重建特征在语义上保持一致，但强度各异，从而有助于在训练数据有限的情况下提升鲁棒性和泛化能力。

3.3 上下文引导特征融合模块

所提出的 CG-Fuse 模块的详细结构如图7(a)所示。具体而言，来自较深解码器层的低分辨率但上下文丰富的特征首先进行上采样（最深层直接上采样，较浅层则通过 CCU 上采样），并通过线性层投影以形成查询（Q）。与此同时，来自高一层 WT-Aug 模块的增强特征被投影为

键（K）和值（V）。随后，将三元组（Q, K, V）输入多头交叉注意力机制（图7(b)）。对于每个注意力头，注意力图的计算方式如下：

$$Cross\text{-}Attention(Q,K,V) = Softmax\left(\frac{QK^T}{\sqrt{d_K}}\right)V, \quad (4)$$

其中，d_K 表示键的维度。经Softmax归一化的注意力权重确保了学习过程的稳定性，并突出最相关的上下文信息（图7(c)）。最后，将所有注意力头的输出进行拼接，并通过一个前馈层。通过加上原始解码器特征来应用残差跳跃连接，从而确保训练稳定并实现有效的特征融合。

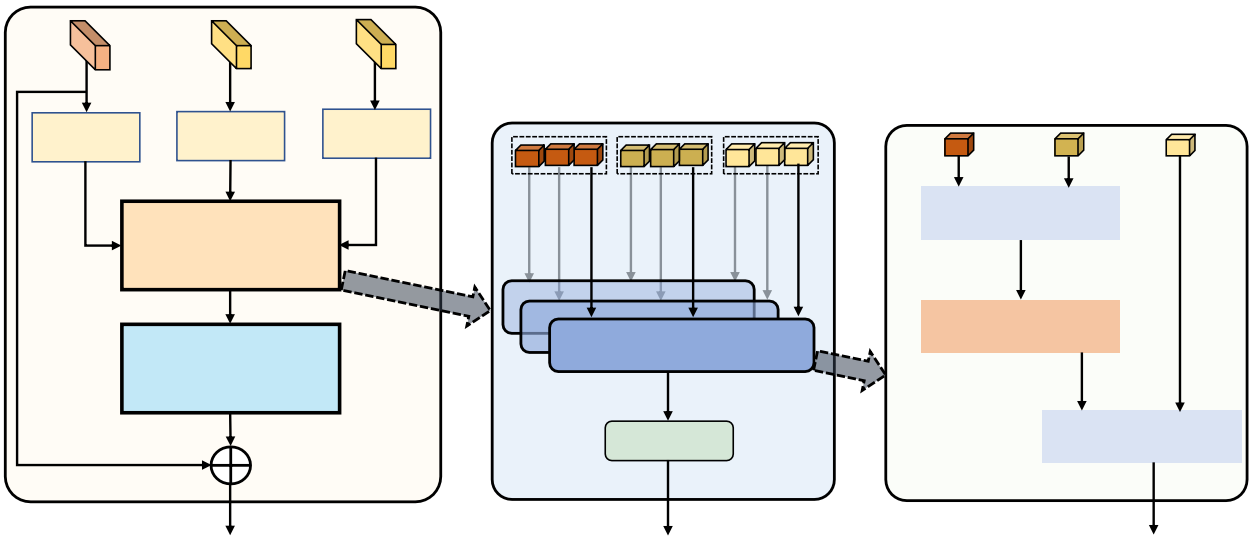


图 7。（a）CG-Fuse、（b）多头交叉注意力和（c）交叉注意力的结构。

4. 实验

4.1 数据集

（1）ACDC数据集

自动心脏诊断挑战赛（ACDC）数据集包含心脏电影MRI扫描，每个病例包含两个时期：收缩末期和舒张末期。所有扫描图像均经过人工标注，标注对象包括 左心室（LV），心肌（MYO）和右心室（RV），作为“真实标注”。按照 [13], 中的数据划分，使用相同的20个病例（40个3D扫描）进行测试，其余80个病例（160个扫描）用于训练。

(2) LA2018 数据集

左心房分割挑战赛（LA2018）数据集 [55] 包含100个3D钆增强磁共振成像扫描（GE-MRIs）及相应的左心房分割掩膜。按照 [56], 中的数据划分，我们使用相同的20个扫描进行测试，其余80个扫描用于训练。

(3) Synapse 数据集

Synapse 多器官分割（Synapse）数据集包含 30 例腹部 CT 扫描，并标注了多个器官，包括主动脉、胆囊、脾脏、左肾、右肾、肝脏、胰腺和胃。按照数据划分方式， [13], 12 例扫描图像留作测试，其余 18 例用于训练。

(4) TN3K 数据集

TN3K 数据集 [57] 是一个公开可用的数据集，旨在支持基于超声图像的甲状腺结节分割研究。该数据集包含来自 2,421 名患者的 3,493 张图像，其中 614 张留作测试，其余 2,879 张用于训练。

(5) Kvasir-SEG 数据集

Kvasir-SEG 数据集 [58] 包含 1,000 张带有相应“真实值”标注的息肉内窥镜图像，此外还包含一个没有公开标签的测试集。在 [58], 该数据集被划分为 880 张训练图像和 120 张验证图像。在本研究中，我们使用这 880 张图像作为训练数据集，并使用 120 张图像进行测试。

(6) ISIC2018 数据集

ISIC2018 数据集 [59] 是国际皮肤影像协作组织（ISIC）_ 挑战赛的一部分，广泛用于皮肤病变分割。该数据集包含 2,594 张皮肤镜图像及其对应的专家标注病变掩膜。按照官方挑战赛的划分，2,594 张图像用于训练和验证，而组织方保留一个独立测试集，且不公开其“真实值”标签。在本研究中，我们将这 2,594 张带标签的图像随机划分为 2,074 张训练图像和 520 张测试图像。

表1总结了本研究所用六个基准数据集的特征，包括各自的成像模态、标注结构，以及每个划分中用于训练和测试的病例/图像数量。请注意，训练数量仅表示每个数据集划分中可用于训练的最大扫描图像/图像数量，这些数量

可用于训练。为模拟少样本学习场景，我们仅从中随机选取少量病例用于模型训练（详见结果部分）。无论少样本训练使用了多少扫描图像/图像，测试集划分中的所有样本均用于模型评估。

表1。用于评估的六个公开数据集概述。

数据集	模态	目标	训练	Test
ACDC	MR	心脏	160	40
LA2018	MR	左心房	80	20
Synapse	CT	腹部器官	18	12
TN3K	US	甲状腺	2,879	614
Kvasir-SEG	内窥镜检查	息肉	880	120
ISIC2018	皮肤镜检查	皮肤病变	2,074	520

4.2 实现细节

为全面评估 DINO-AugSeg 的有效性，我们将其与卷积和 Transformer 两种范式下具有代表性的先进分割模型进行比较。卷积基线包括 Attention-Unet [60], MultiResUNet [61], nnU-Net [11], SegNet [62], SegResNet [63], U-Net [7], 和 Unet++ [10]，而基于 Transformer 的模型包括 MISSFormer [64], SegFormer [48],和 SwinUNETR [15]。我们还将 DINO-AugSeg 与领先的自监督表征学习方法进行基准比较，包括 AIMv2 [65], 掩码自编码器 [21], MoCov2 [66], 和 SimSiam [67], ；这些方法将自监督特征用作 U-Net 风格架构中的冻结编码器。此外，我们还将本方法与近期提出的 SegDINO [49], 进行比较。该方法同样冻结 DINOv3 编码器，但使用轻量级的基于 MLP 的解码器。

所有实验均使用 PyTorch 框架实现，并在运行 Ubuntu 22.04.5 LTS 的 NVIDIA RTX 4090 GPU 上执行。为确保公平比较，所有模型均使用 Adam 进行优化，学习率设为 1×10^{-4}。训练目标采用交叉熵损失和 Dice 损失的加权组合，两者权重相同。对于 SegDINO 和我们提出的 DINO-AugSeg，DINOv3 骨干网络保持冻结，仅训练分割解码器。同样，对于其他自监督对比方法（AIMv2、掩码自编码器、MoCov2 和 SimSiam），特征提取编码器保持冻结；而其他所有模型均采用端到端方式训练。表 2 汇总了各数据集的实验配置，包括输入图像尺寸、批次大小和训练轮数。所有模型均在二维条件下进行训练和测试。对于 DINO-AugSeg，WT-Aug模块仅在训练期间使用，测试阶段不使用。评估时，我们采用 Dice 相似系数（DICE）和第 95 百分位豪斯多夫距离

（HD95）。具体而言，ACDC、LA2018 和 Synapse 的 DICE 和 HD95 在三维体积上计算，而 TN3K、Kvasir-SEG 和 ISIC2018 的指标则在二维图像上计算。对于具有多个分割目标的数据集（例如 ACDC 或 Synapse），相应指标取各分割目标上的平均值。

表 2。 各数据集的实验设置，包括所用的输入图像尺寸、批次大小和训练轮数。

数据集	图像尺寸	批次大小	轮数
ACDC	768×768	2	2000
LA2018	768×768	2	1000
Synapse	224×224	2	500
TN3K	512×512	4	500
Kvasir-SEG	512×512	2	1000
ISIC2018	224×224	4	250

4.3 结果

4.3.1 少样本学习的性能

（1）少样本

ACDC、LA2018 和 Synapse 数据集上的单样本分割结果（体积图像）我们首先在“单样本”训练场景下评估了 DINO-AugSeg 的分割性能，其中每个模型仅使用一张带标注扫描图进行训练。表3总结了三个体积数据集上的结果：ACDC、LA2018 和 Synapse，并将我们的方法与一系列最先进的方法进行了比较，包括基于卷积和基于Transformer的架构，以及近期的自监督学习方法。如结果所示，DINO-AugSeg 在 ACDC 和 LA2018 数据集上取得了最高的 Dice分数和最低的 HD95值，并在 Synapse 上展现出具有竞争力的性能。这些结果表明，即使在极低数据场景下，DINO-AugSeg 仍具备很强的泛化能力和鲁棒性。

表3。 使用一个训练扫描在 ACDC、LA2018 和 Synapse 数据集上的实验结果。箭头指向准确率提升的方向。粗体指标结果表示最佳值。

数据集		ACDC		LA2018		Synapse	
方法	Type	DICE (%)↑	HD95 (体积) ↓	DICE (%)↑	HD95 (体积) ↓	DICE (%)↑	HD95 (体积) ↓
Attention-Unet	Conv	0.73	106.65	23.73	44.48	39.93	87.19
MultiResUNet	Conv	9.41	82.46	21.85	53.70	0.57	196.07
nn-UNet	Conv	22.45	91.13	55.03	40.56	23.30	183.46
SegNet	Conv	22.26	64.35	37.46	29.37	17.09	87.25
SegResNet	Conv	22.84	76.08	29.10	45.07	29.20	121.51
U-Net	Conv	25.89	58.39	20.25	44.83	42.08	97.74

Unet++	Conv	22.86	64.08	24.96	41.25	44.56	108.02
MISSFormer	Trans	23.39	54.24	32.46	40.48	20.15	87.10
SegFormer	Trans	17.16	75.81	36.72	27.04	15.48	90.68
SwinUNETR	Trans	25.61	63.58	35.34	33.62	24.65	129.46
AIMv2	SSL	47.29	40.32	49.25	24.72	33.62	94.36
MAE	SSL	41.63	18.27	47.79	28.17	3.34	118.75
MoCov2	SSL	60.64	44.84	39.04	36.63	44.71	85.17
SimSiam	SSL	49.32	41.31	22.72	32.38	**45.02**	93.29
SegDINO	SSL	60.39	26.39	48.45	34.20	20.88	**68.32**
DINO-AugSeg	SSL	**71.70**	**23.69**	**75.43**	**22.06**	42.84	93.47

我们进一步评估了“七样本”训练场景下的性能，其中使用七张带标注扫描图进行训练。表4展示了相同三个数据集上的结果。同样，DINO-AugSeg 在所有数据集上始终取得最高的 Dice分数，同时保持具有竞争力的 HD95值。这些结果表明，与传统基线和自监督基线相比，DINO-AugSeg 能够有效利用有限的标注数据，实现更优的分割性能。具体而言，与表3中的 h ' - h ' 相比，三个数据集上的 Dice分数分别提高了 10.15%，10.79%，以及 28.35%（绝对值，以下同），此外，在这些数据集上使用两张训练扫描图（' - h '）获得的结果见补充材料，其性能提升趋势相同。

表4。在 ACDC、LA2018 和 Synapse 数据集上使用七张训练扫描图像的实验结果。箭头表示准确率提升的方向。粗体显示的指标结果为最佳值。

数据集		ACDC		LA2018		Synapse	
方法	编码器类型	DICE (%)↑	HD95 (体积)↓	DICE (%)↑	HD95 (体积)↓	DICE (%)↑	HD95 (体积)↓
Attention-Unet	Conv	61.98	30.59	56.48	33.10	69.76	24.44
MultiResUNet	Conv	26.46	55.72	48.53	39.47	5.33	250.27
nn-UNet	Conv	52.09	49.57	81.54	12.58	63.11	30.55
SegNet	Conv	54.75	31.42	61.41	20.90	52.97	29.13
SegResNet	Conv	54.02	52.81	63.04	28.16	67.57	26.64
U-Net	Conv	60.72	27.51	53.45	33.74	66.02	25.03
Unet++	Conv	62.79	34.87	64.47	30.85	70.85	19.47
MISSFormer	Trans	41.04	20.87	76.38	18.04	51.95	34.94
SegFormer	Trans	39.87	43.60	65.11	23.40	43.51	41.11
SwinUNETR	Trans	48.49	24.74	74.73	19.80	53.10	33.55
AIMv2	SSL	61.23	18.75	77.56	16.74	57.23	19.75
MAE	SSL	47.99	15.04	73.40	21.68	48.48	32.19
MoCov2	SSL	73.23	10.35	62.90	22.95	66.91	19.00
SimSiam	SSL	75.07	12.01	57.63	25.88	66.44	**17.60**
SegDINO	SSL	73.96	**6.18**	83.58	12.48	54.10	22.44
DINO-AugSeg	SSL	**81.85**	9.96	**86.22**	**11.75**	**71.19**	21.76

(2) 少样 本分割在 TN3K、Kvasir-SEG 和 ISIC2018 数据集上的结果（二维图像 ）

我们进一步评估了 DINO-AugSeg 在三个二维图像数据集上的少样本分割性能：TN3K、Kvasir-SEG 和 ISIC2018，分别涵盖超声、内窥镜检查和皮肤镜检查模态。在该实验中，分别使用 TN3K、Kvasir-SEG 和 ISIC2018 上的 25/100、10/40 和 5/25 个带标注的 h b f ‘f - h ’ 样本进行训练。表5汇总了 DINO-AugSeg 的 DICE 结果，并与各种基于卷积、基于 Transformer 以及自监督学习的方法进行了比较。

如图所示，在 TN3K（25 个训练样本）和 Kvasir-SEG 的两种少样本设置下，DINO-AugSeg 均取得了最高的 Dice分数，大幅超越传统卷积和基于 Transformer 的基线方法。在 TN3K（100 个训练样本）上，SegDINO 的表现最佳，而 DINO-AugSeg 紧随其后。对于 ISIC2018，尽管 DINO-AugSeg 的表现略逊于 SimSiam 和 MoCov2，但在基于自监督学习的模型中仍取得了具有竞争力的结果。总的来说，这些发现凸显了 DINO-AugSeg 在少样本分割场景下，面对异构成像模态和不同数据规模时展现出的强大适应能力和泛化能力。

表5。 TN3K、Kvasir-SEG 和 ISIC2018 图像数据集上的 DICE 性能，分别使用 25/100、10/40 和 5/25 个训练样本。箭头指向准确率提升的方向。粗体指标结果表示最佳值。

数据集		TN3K		Kvasir-SEG		ISIC2018	
训练样本数量		25	100	10	40	5	25
方法	编码器类型	DICE (%)↑	DICE (%)↑	DICE (%)↑	DICE (%)↑	DICE (%)↑	DICE (%)↑
Attention-Unet	Conv	48.01	45.65	32.78	42.59	57.40	76.98
MultiResUNet	Conv	30.15	37.35	31.85	34.14	61.79	72.06
nn-UNet	Conv	43.40	45.46	38.24	30.08	65.90	70.33
SegNet	Conv	44.18	55.04	32.34	39.09	54.69	71.93
SegResNet	Conv	42.97	51.75	31.10	29.91	62.68	74.12
U-Net	Conv	46.04	45.46	30.84	45.31	56.52	78.94
Unet++	Conv	47.66	55.04	17.34	64.73	53.58	76.73
MISSFormer	Trans	23.09	47.76	16.88	25.56	61.85	75.72
SegFormer	Trans	35.05	45.46	15.41	39.66	67.74	77.53
SwinUNETR	Trans	40.76	46.73	39.92	23.11	62.95	73.04
AIMv2	SSL	53.08	50.56	57.22	54.90	66.73	74.44
MAE	SSL	49.45	52.99	42.69	66.96	69.99	74.06
MoCov2	SSL	56.08	62.78	60.63	63.84	68.60	81.27
SimSiam	SSL	56.09	60.80	57.10	44.94	**73.02**	**81.72**
SegDINO	SSL	30.06	**66.16**	64.38	42.60	64.85	79.38
DINO-AugSeg	SSL	**60.13**	65.40	**73.59**	**78.61**	67.64	78.53

4.3.2 消融实验

(1) 不同增强策略的影响

我们在 ACDC数据集上进行了消融实验，以评估所提出的 DINO-AugSeg框架中不同增强策略的影响。我们研究了三类增强：图像级、特征级（空间域）和特征级（小波域）增强。

对于图像级和特征级（空间域）增强，训练期间会在图像或特征的空间域上随机应用四种常用的强度变换：亮度调整、运动模糊、泊松噪声和随机像素遮蔽。对于特征级（小波域）增强，如方法部分所述，仅在小波域对 DINOv3 特征图应用随机像素遮蔽。这种简化的设计使我们能够单独且清晰地评估小波域增强对最终分割性能的贡献。

如表 6 所示，与图像级增强相比，特征级增强策略通常能够提升分割准确率；就 DICE 而言，小波域增强在不同心脏结构上带来的提升最为稳定。我们观察到，在少样本设置下，无论是在 DINO-AugSeg框架中于空间域还是小波域进行增强，特征级增强的 HD95 都不一定达到最佳表现。我们推测，直接增强 DINOv3 特征图可能会引入额外噪声，使模型更难以精确捕捉精细的目标边界，进而影响 HD95 指标。

表 6。 在 ACDC数据集 上采用不同增强策略的消融实验。

增强	领域	1 训练样本		2 训练样本		7 训练样本		全部训练样本	
		DICE (%)↑	HD95 (体积) ↓	DICE (%)↑	HD95 (体积) ↓	DICE (%)↑	HD95 (体积) ↓	DICE (%)↑	HD95 (体积) ↓
图像级	空间	67.80	**17.95**	68.68	13.86	75.16	**5.80**	86.08	3.21
特征级	空间	68.65	19.47	70.52	**6.71**	80.38	8.53	90.38	1.75
	小波	**71.70**	23.69	**76.24**	15.56	**81.85**	9.96	**91.29**	**1.66**

（2）使用 DINOv3 作为编码器时不同解码器之间的比较

为进一步评估所提出的 CG-Fuse 解码器的有效性，我们将其与六种广泛采用的解码器架构进行了比较：DeepLabv3+ [46]、OCRNet [68], PSPNet [47], SegFormer、U-Net，以及

UNet++，同时保持 DINOv3 作为共享编码器。实验在 ACDC 和 LA2018 数据集上采用七样本训练设置进行，此时仅有七幅带标注扫描图可用于训练。如表7所示，所提出的 CG-Fuse 解码器在两个数据集上均始终取得更优的分割性能。在 ACDC 和 LA2018 数据集上，其 Dice分数分别达到最高的 81.85% 和 86.22%，HD95值分别为 9.96 和 11.75，表现同样出色。这些结果表明，CG-Fuse 解码器能够有效增强多尺度特征融合和空间一致性，在卷积和基于 Transformer 的框架中均优于传统解码器。

表7。在 ACDC 和 LA2018 数据集的七样本训练设置下，对 DINO-AugSeg 采用不同解码器架构的比较。箭头指向准确率提升的方向。加粗的指标结果表示最佳值。

数据集	ACDC		LA2018	
解码器	DICE (%){↑}	HD95（体积）{↓}	DICE（%）↑	HD95（体积）↓
Deeplabv3plus	63.63	9.85	73.79	20.85
OCRNet	73.41	10.08	72.89	12.85
PSPNet	70.93	**5.07**	71.98	13.97
SegFormer	77.68	6.98	81.51	15.89
U-Net	72.81	10.24	84.62	13.12
UNet++	79.34	6.71	84.13	15.55
CG-Fuse（本文方法）	**81.85**	9.96	**86.22**	**11.75**

（3）七样本训练设置下 WT-Aug 和 CG-Fuse 的效果

为进一步探究 WT-Aug模块和所提出的CG-Fuse解码器的作用，我们在ACDC和LA2018 数据集上进行了七样本训练设置下的消融实验。在该实验中，我们比较了四种配置：使用和不使用 WT-Aug，以及使用和不使用CG-Fuse（不使用时则替换为标准的类U-Net解码器，直接拼接编码器特征和解码器特征）。

如表8所示，WT-Aug 和 CG-Fuse 单独使用时均提升了分割性能，而二者结合则取得了最佳的整体结果。具体而言，在 ACDC 和 LA2018 数据集上，同时采用这两个模块分别取得了最高的 Dice分数 81.85% 和 86.22%，以及最低的 HD95值 9.96 和 11.75。这些发现证实，WT-Aug 提升了数据和特征的多样性，而 CG-Fuse 则增强了解码阶段的特征交互和空间一致性。

表8。在 ACDC 和 LA2018 数据集的七样本训练设置下，WT-Aug 和 CG-Fuse 对 DINO-AugSeg 的影响。箭头表示准确率提升的方向。粗体标示的指标结果为最佳值。

数据集		ACDC		LA2018	
WT-Aug	CG-Fuse	DICE (%) ↑	HD95（体积）↓	DICE（%）↑	HD95（体积）↓
✓	✓	**81.85**	**9.96**	**86.22**	**11.75**
×	✓	76.98	16.87	85.61	13.79
✓	×	77.32	11.82	85.05	16.23
×	×	72.81	10.24	84.62	13.12

（4）. 图像尺寸对分割性能的影响

我们观察到，DINO-AugSeg 并未在所有数据集上都取得最先进性能，尤其是在 Synapse/ISIC2018 上，SimSiam 和 MoCov2 等几种基于自监督学习的方法表现更好。我们推测，这一局限可能源于 DINOv3 中经过下采样的输入表征；DINOv3 最初是在大矩阵尺寸的高分辨率图像上训练的。然而，对于 Synapse/ISIC2018 这类小图像尺寸的数据集（表2），过度下采样输入会导致精细空间细节丢失，而这些细节对于精确的边界勾画至关重要。为验证这一假设，我们进行了另一项消融实验，将 Synapse/ISIC2018 数据集上采样至更高分辨率（512×512），以抵消 DINOv3 编码器的下采样效应。

表9。 在不同训练数据量下（Synapse 的 1、2 和 7 张扫描图像；ISIC2018 的 5、10 和 25 张图像），不同输入图像尺寸对 SimSiam 和 DINO-AugSeg 性能的影响，以 Dice分数表示。

方法	数据集	Synapse			ISIC2018		
Name	图像尺寸	1	2	7	5	10	25
SimSiam	224×224	45.02	47.00	66.44	73.02	78.12	81.72
SimSiam	512×512	50.11	55.60	69.56	72.35	73.90	77.70
DINO-AugSeg	224×224	42.84	50.37	71.19	67.64	74.09	78.53
DINO-AugSeg	512×512	**54.65**	**63.92**	**77.24**	**77.40**	**81.14**	**83.81**

表9总结了不同输入图像尺寸对 SimSiam 和 DINO- 的性能影响

在不同训练样本数量下的 AugSeg（Synapse 使用 1、2 和 7 张扫描图像，ISIC2018 使用 5、10 和 25 张图像）。如图所示，在 Synapse 上，将输入分辨率从 224×224 增至 512×512 ，两种方法的 Dice分数均持续提升，其中 DINO-AugSeg 的性能增幅尤为显著。对于 ISIC2018，我们观察到，DINO-AugSeg 在较高分辨率下取得了显著提升，且优于 SimSiam。这表明，使用更大的输入矩阵可以减轻 DINO-AugSeg 中 DINOv3编码器的下采样效应，从而有助于保留细粒度细节，实现精准的图像分割。相比之下，SimSiam 在 ISIC2018

数据集上将输入尺寸从 224×224 增至 512×512时，性能有所下降。这可能是因为 SimSiam 使用了在矩阵尺寸为 224×224的自然图像上预训练的 ResNet-50 编码器，因此原始的 224x224 输入尺寸更为合适。然而，对于基于计算机断层扫描的 Synapse 数据集，SimSiam 在较大的输入尺寸下也有所提升，这表明医学图像特征与矩阵尺寸之间存在复杂的相互作用。

4.3.3 视觉比较

为定性比较分割性能，图 8 展示了在 7样本训练设置下，ACDC、LA2018 和 Synapse 数据集在磁共振（第一、二行）和计算机断层扫描（第三行）模态上的代表性结果。第一列展示叠加在原始图像上的真实标注掩膜，其余各列则展示 DINO-AugSeg、nn-UNet、SwinUNETR、SimSiam 和 SegDINO 生成的预测叠加图，分别代表先进的基于卷积的方法（nn-UNet）、基于 Transformer 的方法（ SwinUNETR ）以及自监督学习方法（ DINO-AugSeg 、 SimSiam 和 SegDINO）。可以看出，DINO-AugSeg 在磁共振和计算机断层扫描图像的多种解剖结构上均表现出稳健且稳定的分割性能。

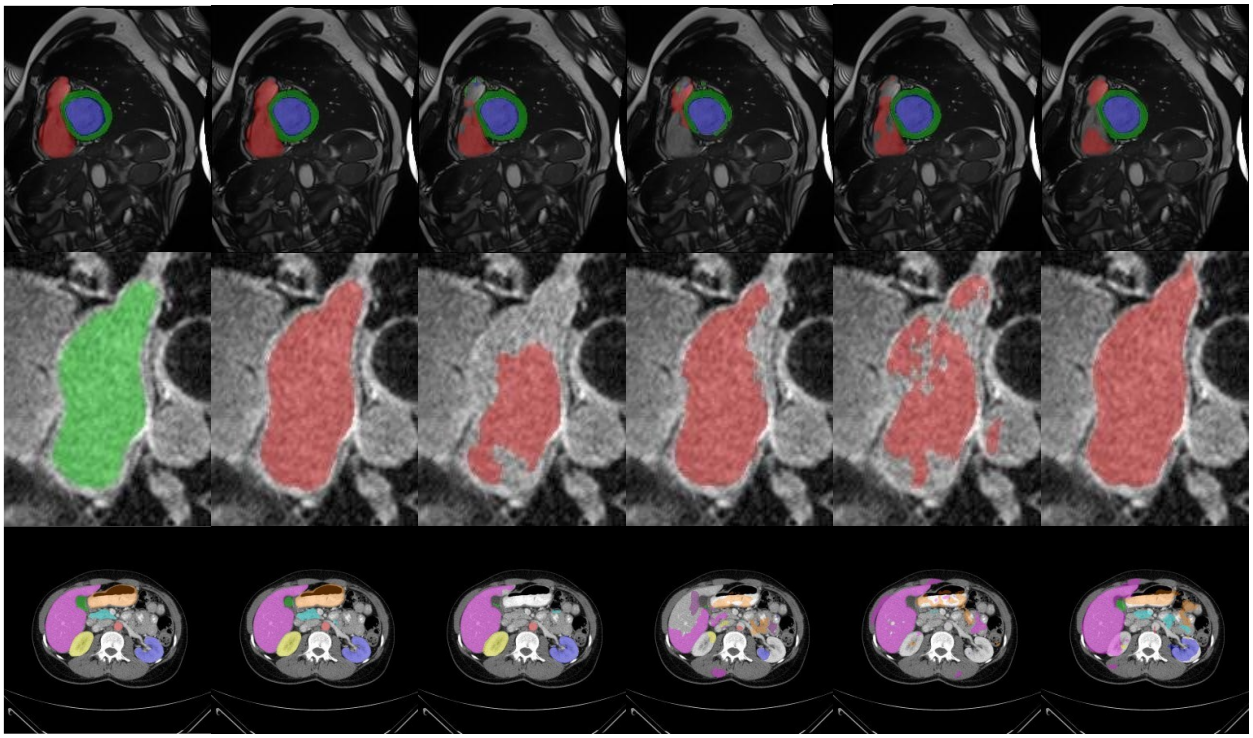


图 8. ACDC、LA2018 和 Synapse 数据集在磁共振（第一、二行）和计算机断层扫描（第三行）模态下具有代表性的定性分割结果。第一列显示了真实标注——

真实标注叠加图，其余各列则展示不同方法在 7样本训练设置下的预测结果。

图9进一步展示了在TN3K、Kvasir-SEG和ISIC2018 数据集（从上到下）上，分别采用25样本、10样本和5样本训练设置时的视觉比较，呈现了所提出的DINO-AugSeg的代表性分割结果。第一列展示了叠加在原始图像上的真实标注掩膜（绿色），其余各列则展示了DINO-AugSeg、nn-UNet、SwinUNETR、SimSiam和SegDINO生成的预测掩膜（蓝色）。在这三个数据集上，DINO-AugSeg在少样本学习场景下始终取得更优的分割性能。补充材料中提供了这五个多模态数据集在各种少样本设置下的更多定性示例。

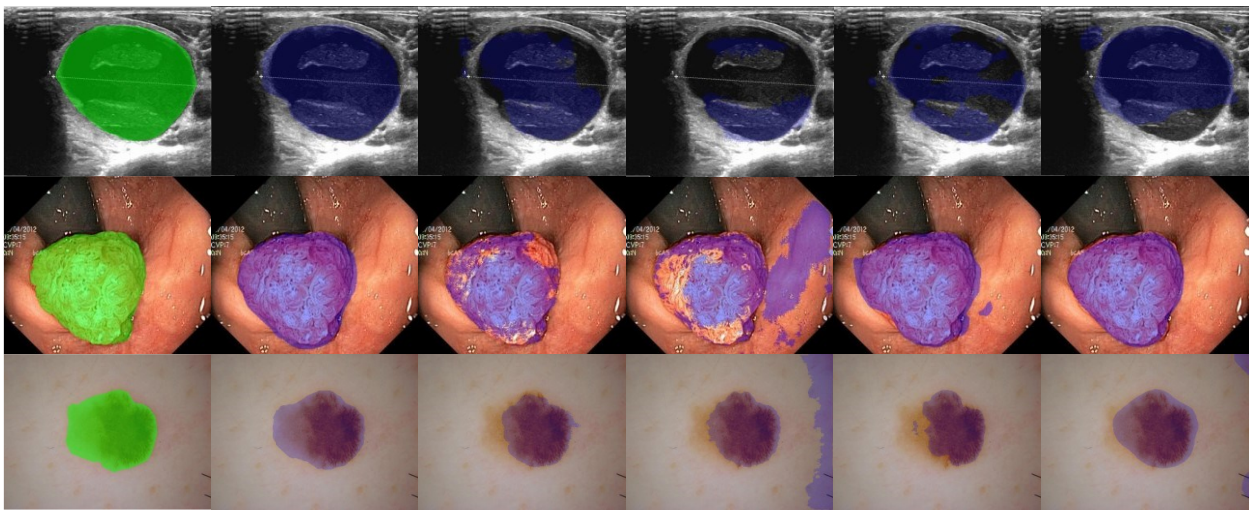


图9。 在超声（US）、内窥镜检查和皮肤镜检查模态下，TN3K、Kvasir-SEG和ISIC2018 数据集分别采用25样本、10样本和5样本训练设置时的代表性分割结果。真实标注掩膜以绿色显示，模型预测结果以蓝色显示。

5. 讨论

本研究利用基于DINOv3的特征进行少样本医学图像分割，重点关注特征级增强和上下文引导融合。在六个数据集上的实验结果表明，所提出的DINO-AugSeg框架具有良好效果。本节将重点阐述关键见解，并讨论其局限性，以此指出未来研究的潜在方向。

数据增强是一种广泛采用的策略，可增加训练样本的多样性。由于标注数据有限，这在医学图像分割中尤为重要。

与自然图像不同，医学图像通常受益于融入先验知识的领域特定增强策略，例如基于图谱的 [27] 和基于形态学的方法 [69]，这些策略有助于提升模型性能。在这项工作中，我们提出了一种针对 DINOv3 特征的小波域特征级增强方法。尽管该方法在少样本分割中效果显著，但与全样本训练相比，仍存在明显的性能差距。将领域先验作为约束融入基于小波的增强中，或有助于进一步提升性能。例如，鉴于 DINOv3 特征已展现出对象识别和关键点检测能力，以对象为中心的特征增强可能尤其有益。

与此同时，我们的结果表明，小波域特征级增强在精确边界勾画方面可能面临一些挑战，尤其是在少样本学习场景下，这一点可由 HD95 指标反映出来。这表明，在小波域中采用形状感知或结构保持增强，或许有助于引导模型更好地捕捉对象轮廓的细节。未来，我们计划探索将与对象和形状相关的先验融入小波域特征增强，以进一步提升少样本医学图像分割的鲁棒性和边界准确度。

此外，解码器通过上采样逐步提高特征图的空间分辨率，在整合关键特征和抑制无关信息方面发挥着至关重要的作用。大多数现有方法侧重于多尺度特征融合，或为不同通道或空间区域分配重要性权重。本研究表明，在少样本学习场景中，利用预训练的 DINOv3 特征通过交叉注意力引导解码器融合是有效的。然而，与 nn-UNet 和 SegFormer 等最先进的方法相比，随着训练样本数量增加，这一优势会逐渐减弱。这一现象表明，在自然图像上训练的 DINOv3 特征可能并非完全适用于医学图像分析，这反映了自然图像与医学图像之间已知的领域差距。一个有前景的未来研究方向是探索先进的微调策略，例如适配器或低秩适配（LoRA）[8], ，以便让 DINOv3 特征更好地适应医学领域。此外，将所提出的 DINO-AugSeg框架扩展至磁共振成像或计算机断层扫描数据中的三维对象分割，也是一个有价值的方向，因为当前设计仅限于基于二维切片的分割，且未明确利用切片间空间依赖关系。

6. 结论

本研究探讨了 DINOv3 特征在少样本医学图像分割中的有效性。针对训练样本有限所带来的挑战，我们提出了 WT-Aug（一种小波域特征级增强策略）和 CG-Fuse（一种用于解码器的上下文信息引导的特征融合模块）。在涵盖五种医学成像模态的六个公开基准上开展的大量实验表明，我们的 DINO-AugSeg框架在少样本设置下表现出色。这些结果凸显了 WT-Aug 和 CG-Fuse 在增强表征学习和提高分割准确率方面的有效性，表明它们有望成为推进少样本医学图像分割的通用策略。

致谢

本研究得到了美国国立卫生研究院的支持（R01 CA240808、R01 CA258987、R01 EB034691 和 R01 CA280135）。

参考文献

X, ., , ., U p, J. ., T, Y., T i i, D. .：《‘Di N：p i c v i c c chi c f y ph i ET T i’》，p iz ic I i phic，0，，第 0 5 页X, ., i, X., h, . ., L, J., L, W., Zh, Y.：《‘y hi-i ch-b i-p vi i f k f ic i i’》，hy，0 5i h, ., J, ., i, ., i, ., D i, J., y, ., y, L.：《‘D p i f i i i i i h py p i：vi’》，J f ic I i R i i O c y，0，5，（5），第 57 595 页bi, .R., B, ., , ., L, ., , ., Ri, . ., Ni, T. ., i chb, J., i z, J., W i z, J.：《‘ifici I i c f c ic i c i c p b i c ic p c：p y f ibi i y y’》，E p J f ic O c y，0，50，（），第 0 99 5 页X, ., U p, J. ., T, Y., O h, D., , ., T i i, D. .：《‘R-LN-D：ic y c i i b i q ific i i h cic y ph z vi D ET T i i h i i’》，hy，0 0，7，（），第 7R b, O., i ch, ., B, T.：《‘U：v i k f bi ic i i’》，i E i (E .)^(E .)：《‘B k U：v i k f bi ic i i’》（p i，0 5，.），第 7 页z, R., h, E. ., R, ., Ji, Y., vv, . ., B z p, ., i i f bi, ., h, J. ., i, E., h f, D.：《‘ic i i vi：Th cc f’》，IEEE T c i y i chi I i cX, ., U p, J. ., L, J., Zh, ., Y, Y., R y, .B., , ., Ni, L., R hi, Y., Li, W.：《I h ic i i p b v? v y f c v p f》

[无法准确翻译：原文字符损坏严重，无法辨识]

IEEE 计算机学会计算机视觉与模式识别会议论文集，0，7，（），第 5557 页 Zh，.，B k i h，.，D，.，J. .，D c，..：《D i i f i f h ic i i》，i E i（E .）^(E .)：《B k D i i f i f h ic i i》（0 9，.），第 5 55 页，L.v. .，i，.：《i izi i NE》，J f chi i ch，00，9，（N v），第 579 05 9 页 B ch，E.，cI，L.，y，J.，D，. .，k，I.W.，N，L. .，i h，.，N，E.W.：《Di i i y c i f vi izi i c i U》，N bi ch y，0 9，7，（），第 0 页 D c，T.，Oq b，.，i，J.，B ki，.：《i i f i》，Xiv 预印本 Xiv: 09.5，0 i，.，h z，N.，，N.，U zk i，J.，J，L.，z，.N.，i，Ł.，khi，I.：《i i y》，v c i i f i p c i y，0 7，0 h，.，h h f，T. .：《v y i i f p i》，J f bi，0 9，，（），第 D i，T.，T y，.W.：《I p v iz i f c v i k i h c》，Xiv 预印本 Xiv: 70.0 55，0 7 Zh，Z.，Zh，L.，，.，Li，.，Y，Y.：《R i i》，i E i（E .）^(E .)：《B k R i i》（0 0，.），第 00 00 5 页 ich h f，.，Li，Y.，，.：《k c p i p》，v c i i f i p c i y，0，5，第 59 595 页 T，Z.，，Y.，W，J.，W，L.：《i：k c ffici f f p vi vi p i i》，v c i i f i p c i y，0，5，第 007 009 7 页 W，L.，，B.，Zh，Z.，T，Z.，，Y.，W，Y.，W，Y.，i，Y.：《i v：c i vi k c i h ki》，i E i（E .）^(E .)：《B k i v：c i vi k c i h ki》（0，5 9 5 0 .），第 Li，Y.，L，T.，W，Y.，，.，h，Y.，Y，X.：《k：E h ci ic i i i h c c i i f ki》，i E i（E .）^(E .)：《B k k：E h ci ic i i i h c c i i f ki》（p i，0，.），第 9 h，Z.，，D.，，.，f，W.，B，. .，B，.：《k i i v c ic i y i》，i E i（E .）^(E .)：《B k k i i v c i c i y i》（0，.），第 970 9 0 0 Xi，Z.，Zh，L.，Y，L.，Xi，Z.，W，L.：《yb i k i i f ic i i》，IEEE 生物医学与健康信息学汇刊，0，，（），第 5 5 h，. .，h y，.：《i i c v i k i h i i i i》，i E i（E .）^(E .)：《B k i i c v i k i h i i i i》（IEEE，0 7，.），第 590 59 页 Zh，X.，，Y.，Li，.，L，J.，i，L.，，J.，Y，L.：《f i f i i z i i i c i i》，i E i（E .）^(E .)：《B k f i f i i z i i i c i i》（IEEE，0，9 9 .），第 Y，L.，i，L.，，L.，Zh，W.，hi，Y.：《R v i i i k c i cy i i p vi i c i》，i E i（E .）^(E .)：《B k R v i i i k c i cy i i p vi i c i》（0，.），第 7 7 Y，L.，Zh，Z.，Zh，.：《U i ch v：hi h i i f i p vi i c

i》，IEEE 计算机视觉与模式识别汇刊，0 5 5 iv v，N.，i，.，izh v ky，.，k v，I.，kh i v，R.：《D p：i p y p v k f v fi i》，Th f chi i ch，0，5，（），第 9 9 95 页 h，L. .，Zh，Y.，p，.，ch ff，.，，.：《E c c i h p b c v i fic i i》，i E i（E .）^(E .)：《B k E c c i h p b c v i fic i i》（
018, edn 01-818 .），第 47 ^
Zh，.，hi，J.，i，X.，W，X.，Ji，J.：《y i c p i k》，i E i（E .）（E .）：《B k y i c p i k》（0 7，.），第 90 Xi，E.，W，W.，Y，Z.，k，.，v z，J. .，L，.：《：i p ffici i fic i i h f》，v c i i fi p c i y，0，，第 077 090 9 页 Y，，W，.，Xi，Z.，h，.，Zh，L.：《i：ffici i fic i i i h i v》，Xiv 预印本 Xiv: 509 00，0 5 50 . W，Y.，Li，X.，Li，J.，Y，.，Zh，.，Zh，.：《Di i ic i：E p i i c ffi i y f k y p vi ic i》，i E i（E .）^(E .)：《B k Di i ic i：E p i i c ffi i y f k y p vi ic i》（0，.），第 9 97 5 yz b，L.，i y，R.，p，.：《Di v b f p vi i f f h ic i i》，i E i（E .）^(E .)：《B k Di v b f p vi i f f h ic i i》（IEEE，0，.），第 5 5 i h，.，h kk p i，R.，h h i，.，h，L.，h，.，，J.，，.，i，J.：《hif i chi p vi i：i ffici i f p vi i f ic ic i i c ific i》，ci ific R p，0，，0 0 5，（），第 Y.，Li，.，Y，.，W，X.，，X.：《Di：E p i i hi h fi i y f f f i f ic i i》，Xiv 预印本 Xiv: 50 . 0909，0 5 5 X，.，Li，W.，Zh，X.，Li，.，，X.，W，X.：《v p i：i p b ff c iv p i f ic i》，c i i，0，，第 09 9 55 Xi，Z.，Xi，.，，Z.，，N.，Bi，.，Zh，Y.，，.，R vik，N.，i，.，Y，X.：《b b ch k f i h f i h f i f i i i h c c i c ic c i》，ic i y i，0，7，第 0 5 Y，L.，W，.，Li，X.，. W.，. .：《U c i y f b i f i p vi D f i i》，i E i（E .）^(E .)：《B k U c i y f b i f i p vi D f i i》（p i，0 9，.），第 05 57 页，，.，h，J.，h，.，Li，.，Li，.，h，.：《Thy i i p i i i f i f hy i》，p i bi y ici，0，55，第 0 9 5 Jh，D.，，. .，Ri，. .，J h，D.，D L，T.，v，.，J h，.D.：《R++：v c chi c f ic i i》，i E i（E .）^(E .)：《B k R++：v c chi c f ic i i》（IEEE，0 9，.），第 5 55 59 页，N.，R b，.，T ch，.，bi，.E.，D z，.，，D.，b，B.，，.，Li py i，.，ch i，.：《ki i y i c i 0：ch h by h i i ki i i c b i（i ic）》，Xiv 预印本 Xiv: 90 .0，0 9 0 ch p，J.，Ok y，O.，ch p，.，i ich，.，i z，B.，ck，B.，R ck，D.：《i k：L i v i i i i c i》，ic i y i，0 9，5，第 97 07 Ib h z，N.，R h，. .：《iR UN：R hi ki h U N chi c f i bi i c i》，N k，0 0，，第 7 7 B i y，，.，，.，ip，R.：《：p c v i c c

Chi C. F. I. I. ’, IEEE C. I. P. Y. I. Chi I. I. C., 0 7, 9, 95 ( ), pp. Y. K., .: ‘D RI B. I. I. I. C. Iz. I.’, I E I (E .) ^ 019, edn 0
(E .): ‘B K D RI B. I. I. I. C. Iz. I.’ (P. I., .), pp. , X., D., Z., Li, D., Y., X., , Y.: ‘I. F.: Ff. C. Iv. F. F. Ic. I. I.’, IEEE T. C. I. Ic. I., 0 , , 9 5 5( ), pp. I. I., E., H. K., ., Li, X., D. F., ., I., ., I., D., I. H., ., , . .T., Bé H., L., , Z.: ‘I. Iv. P. I. I. F. Vi. I. C.’, I E I (E .)^(E .): ‘B K I. Iv. P. I. I. F. Vi. I. C.’ (0 5, .), pp. 9 9 5 H., X., , ., I. H., R., , .: ‘I. P. V. B. I. I. H. C. Iv. I.’, Xiv P. P. I. Xiv: 00 .0 97, 0 0 7 H., X., , .: ‘E. P. I. I. P. I. P. I. I.’, I E I (E .)^(E .): ‘B K E. P. I. I. P. I. P. I. I.’ (0 , 5750 575 .), pp. Y., Y., H., X., W., J.: ‘Ob. C. C. P. I. F. Ic. I.’, I E I (E .)^(E .): ‘B K Ob. C. C. P. I. F. Ic. I.’ (P. I., 0 0, .), pp. 7 90 9 X., ., D. I., Y., Zh., ., Zh., Y., D., J., L., W., Zh., Y.: ‘ : I. K. B. I. F. T. I. I. P. Iv. R. I. I. Th. Py. U. I. Y. Chi.’, Xiv P. P. I. Xiv: 507. 9 , 0 5

补充材料：利用基于DINOv3的自监督特征实现鲁棒的少样本医学图像分割

1. 使用两张训练扫描在 ACDC、LA2018 和 Synapse 数据集（体积图像）上的结果我们仅使用两张训练扫描，在 ACDC、LA2018 和 Synapse 数据集上开展实验，以评估所提出的DINO-AugSeg框架在极其有限监督条件下的有效性。表1展示了涵盖广泛分割模型的 DICE 和 HD95 指标，包括基于卷积的方法（Conv）、基于Transformer的架构（Trans）以及自监督学习方法（SSL）。

总体而言，DINO-AugSeg在ACDC（76.24）和LA2018（81.01）上取得最高的Dice分数，并在Synapse（50.37）上表现出竞争力，而MoCov2（51.91）取得了最高的Dice分数。在边界准确度（HD95）方面，DINO-AugSeg在ACDC（15.56）和LA2018（23.57）上取得了很低的HD95，而在Synapse上，其HD95（37.25）排名第二（AIMv2：36.93）。

表1。在 ACDC、LA2018 和 Synapse 数据集（体积图像）上使用两个训练扫描的实验结果。箭头表示准确率提升的方向。加粗显示的指标结果

表示最佳值。

数据集 / 方法	Type	ACDC		LA2018		Synapse	
		DICE (%)↑	HD95 (体积)↓	DICE (%)↑	HD95 (体积)↓	DICE (%)↑	HD95 (体积)↓
Attention-Unet	Conv	40.10	72.50	50.51	45.16	49.94	64.08
MultiResUNet	Conv	7.64	105.56	34.44	49.79	5.410	221.41
nn-UNet	Conv	27.47	40.09	71.39	24.89	40.24	108.11
SegNet	Conv	30.04	76.89	61.14	32.20	24.10	55.65
SegResNet	Conv	19.17	51.64	53.04	41.25	38.09	74.09
U-Net	Conv	31.93	47.45	51.86	38.41	46.35	44.35
Unet++	Conv	32.77	60.33	45.44	39.64	49.82	41.73
MISSFormer	Trans	33.46	26.45	66.38	20.68	27.16	66.28
SegFormer	Trans	11.62	66.39	52.95	29.70	22.80	72.55
SwinUNETR	Trans	32.17	56.37	54.25	33.48	29.47	60.36
AIMv2	SSL	51.52	25.93	61.09	19.64	38.77	**36.93**
MAE	SSL	36.91	16.49	57.50	27.41	32.65	63.82
MoCov2	SSL	58.62	31.33	66.38	**20.68**	**51.91**	39.89
SimSiam	SSL	53.06	32.00	49.01	36.08	47.00	46.72
SegDINO	SSL	57.13	28.63	62.24	30.53	38.30	41.38
DINO-AugSeg	SSL	**76.24**	**15.56**	**81.01**	23.57	50.37	37.25

2. 使用完整训练扫描在 ACDC、LA2018 和 Kvasir-SEG 数据集上的结果　s

我们进一步在全数据设置下评估了所有竞争方法，此时使用所有可用的训练扫描。此实验旨在评估数据稀缺不再构成限制因素时各方法的性能上限。

表 2 总结了 ACDC、LA2018 和 Kvasir-SEG 数据集上的结果。Attention-UNet、SegNet 和 SegResNet 等经典的基于卷积的模型在 ACDC 和 LA2018 上表现稳健，但在难度更高的 Kvasir-SEG 数据集上表现不佳，尤其是边界准确度方面，其 HD95 值较差。基于 Transformer 的模型（MISSFormer、SegFormer 和 SwinUNETR）在 ACDC 和 LA2018 上取得了有竞争力的 Dice 分数，但在 Kvasir-SEG 上的表现参差不齐。自监督学习（SSL）方法取得了显著提升，其中 SegDINO 在各数据集上均取得了出色的整体表现。我们提出的 DINO-AugSeg 在 ACDC（91.29%）、LA2018（90.88%）和 Kvasir-SEG（86.18%）上均取得了最高的 Dice 分数。值得注意的是，该方法在 Kvasir-SEG 上取得了最低的 HD95（59.96），且优势显著，表明其在复杂内窥镜场景中具有更优的边界保真度。

这些结果表明，即使训练数据充足，DINO-AugSeg 也能有效利用增强驱动的表征学习来提升分割性能。它在 DICE 和 HD95 指标上均表现出稳定的提升，凸显了其在不同成像模态下出色的泛化能力和鲁棒性。然而，将这些全数据结果与少样本设置下的结果进行比较时，我们发现 DINO-AugSeg 的优势不再那么明显。具体而言，在部分数据集上，在全量训练方案下，其相较于 nn-UNet 和 SegDINO 等强基线方法的提升较为有限。这表明，DINO-AugSeg 的主要优势在于数据有限时能够提升表征质量；而在数据充足的情况下，各方法之间的性能差距会缩小。

表 2. 在 ACDC、LA2018 和 Kvasir-SEG 上使用全部训练样本的实验结果。箭头表示准确率提升的方向。粗体标出的指标结果为最佳值。

数据集		ACDC		LA2018		Kvasir-SEG	
方法	Type	DICE (%)↑	HD95 (体积) ↓	DICE (%)↑	HD95 (体积)↓	DICE (%)↑	HD95 (体积)↓
Attention-Unet	Conv	91.04	2.47	71.15	28.49	77.65	124.13
MultiResUNet	Conv	60.88	37.19	50.91	48.31	31.50	221.85
nn-UNet	Conv	91.16	1.65	90.64	**5.39**	72.92	114.35
SegNet	Conv	90.14	4.57	86.29	17.53	66.93	125.70
SegResNet	Conv	89.88	1.82	84.15	20.20	78.04	123.71
U-Net	Conv	88.81	12.88	79.37	18.00	75.10	134.76
Unet++	Conv	90.74	4.82	83.20	18.37	73.63	139.57

MISSFormer	Trans	88.57	**1.56**	90.35	6.64	75.76	105.99
SegFormer	Trans	88.86	6.64	89.01	12.04	65.85	117.26
SwinUNETR	Trans	88.37	2.86	89.55	11.53	71.80	108.45
AIMv2	SSL	76.96	4.11	87.86	12.37	72.79	107.00
MAE	SSL	75.69	7.55	78.62	19.25	64.20	163.17
MoCov2	SSL	90.62	3.13	78.84	23.72	81.81	110.19
SimSiam	SSL	90.13	2.35	75.62	21.28	79.89	125.49
SegDINO	SSL	91.04	1.72	90.48	5.98	73.57	98.20
DINO-AugSeg	SSL	**91.29**	1.66	**90.88**	6.32	**86.18**	**59.96**

3. 模型规模对 DINO-AugSeg 的影响

我们进一步考察了模型规模对所提出的 DINO-AugSeg框架的影响，在 7样本训练设置下评估了三种配置：小型、基础和大型。我们在 ACDC 和 LA2018 数据集上开展实验，以评估提高模型容量能否改善低数据场景下的分割性能。

如表3所示，随着模型规模增大，分割性能持续提升。大型模型取得了最佳的总体结果，在 ACDC 和 LA2018 上的 Dice分数分别为 81.85% 和 86.22%，HD95值分别为 9.96 和 11.75。这些提升表明，扩大 DINOv3 骨干网络的规模能够使模型更好地捕捉细微的解剖结构和复杂的空间变化，最终增强其在不同数据集上的鲁棒性。

表3。在 ACDC 和 LA2018 数据集上，采用 7样本训练设置时，不同模型规模（小型、基础、大型）的 DINO-AugSeg 分割性能。箭头指示准确率提升的方向。

数据集 模型规模	ACDC		LA2018	
	DICE(%) ↑	HD95(体积)↓	DICE(%)↑	HD95(体积)↓
小型	77.59	19.33	81.73	24.02
Base	80.72	10.22	83.75	20.67
大型	81.85	9.96	86.22	11.75

4. 统计分析

为进一步评估所比较方法之间性能差异的显著性，我们采用Wilcoxon符号秩检验进行了成对统计检验。具体而言，计算了

各基线方法与所提出的DINO-AugSeg模型之间的p值。该非参数检验用于评估DINO-AugSeg带来的性能提升在所有测试样本中是否具有统计显著性。我们针对每个分割目标的Dice分数应用了Wilcoxon检验。

如表4所示，在单样本和七样本设置下，针对各解剖结构（RV= 右心室、MYO= 心肌、LV= 左心室），DINO-AugSeg相较于大多数竞争方法均取得了具有统计显著性的改进，p值均低于0.05。唯一的例外是SegDINO、SimSiam和MoCoV2，它们在一个或两个结构上的p值高于0.05。

这一观察表明，这三种方法——均基于自监督预训练编码器——可能会产生与 DINOv3 学得的特征表示更为相似的表示。因此，在某些情况下，这些基于自监督学习的方法与 DINO-AugSeg 之间的性能差异较小，可能未达到统计显著性。表5所示的四个额外数据集（ISIC2018、Kvasir-SEG、TN3K 和 LA2018）的结果进一步支持了这一观察。在这些数据集上，AIMv2、MoCov2 和 SegDINO 等基于自监督学习的方法在若干设置下也呈现出大于 0.05 的 p 值。

表4。在ACDC数据集的单样本和七样本训练设置下评估p值的Wilcoxon符号秩统计检验。

方法	1个训练样本			7个训练样本		
目标	RV	MYO	LV	RV	MYO	LV
Attention-Unet	< 0.05	< 0.05	< 0.05	< 0.05	< 0.05	< 0.05
MultiResUNet	< 0.05	< 0.05	< 0.05	< 0.05	< 0.05	< 0.05
nn-UNet	< 0.05	< 0.05	< 0.05	< 0.05	< 0.05	< 0.05
SegNet	< 0.05	< 0.05	< 0.05	< 0.05	< 0.05	< 0.05
SegResNet	< 0.05	< 0.05	< 0.05	< 0.05	< 0.05	< 0.05
U-Net	< 0.05	< 0.05	< 0.05	< 0.05	< 0.05	< 0.05
U-Net++	< 0.05	< 0.05	< 0.05	< 0.05	< 0.05	< 0.05
MISSFormer	< 0.05	< 0.05	< 0.05	< 0.05	< 0.05	< 0.05
SegFormer	< 0.05	< 0.05	< 0.05	< 0.05	< 0.05	< 0.05
SwinUNETR	< 0.05	< 0.05	< 0.05	< 0.05	< 0.05	< 0.05
AIMv2	< 0.05	< 0.05	< 0.05	< 0.05	< 0.05	< 0.05
MAE	< 0.05	< 0.05	< 0.05	< 0.05	< 0.05	< 0.05
MoCov2	< 0.05	< 0.05	< 0.05	< 0.05	< 0.05	**0.09**
SimSiam	< 0.05	< 0.05	< 0.05	< 0.05	< 0.05	**0.65**
SegDINO	**0.07**	< 0.05	< 0.05	< 0.05	< 0.05	**0.31**

表5. 在少样本训练设置下，针对 ISIC2018、Kvasir-SEG、TN3K 和 LA2018 数据集进行 Wilcoxon符号秩统计检验，评估其 p值。

数据集	ISIC2018		Kvasir-SEG		TN3K		LA2018	
方法	5 训练样本	25 训练样本	10 训练样本	40 训练样本	25 训练样本	100 训练样本	1 训练样本	7 训练样本
注意力-Unet	< 0.05	0.06	< 0.05	< 0.05	< 0.05	< 0.05	< 0.05	< 0.05
MultiResUNet	< 0.05	< 0.05	< 0.05	< 0.05	< 0.05	< 0.05	< 0.05	< 0.05
nn-UNet	< 0.05	< 0.05	< 0.05	< 0.05	< 0.05	< 0.05	< 0.05	< 0.05
SegNet	< 0.05	< 0.05	< 0.05	< 0.05	< 0.05	< 0.05	< 0.05	< 0.05
SegResNet	< 0.05	0.06	< 0.05	< 0.05	< 0.05	< 0.05	< 0.05	< 0.05
U-Net	< 0.05	< 0.05	< 0.05	< 0.05	< 0.05	< 0.05	< 0.05	< 0.05
Unet++	< 0.05	0.71	< 0.05	< 0.05	< 0.05	< 0.05	< 0.05	< 0.05
MISSFormer	< 0.05	0.99	< 0.05	< 0.05	< 0.05	< 0.05	< 0.05	< 0.05
SegFormer	< 0.05	< 0.05	< 0.05	< 0.05	< 0.05	< 0.05	< 0.05	< 0.05
SwinUNETR	< 0.05	< 0.05	< 0.05	< 0.05	< 0.05	< 0.05	< 0.05	< 0.05
AIMv2	**0.26**	< 0.05	< 0.05	< 0.05	< 0.05	< 0.05	< 0.05	< 0.05
MAE	< 0.05	< 0.05	< 0.05	< 0.05	< 0.05	< 0.05	< 0.05	< 0.05
MoCov2	**0.07**	< 0.05	< 0.05	< 0.05	< 0.05	< 0.05	< 0.05	< 0.05
SimSiam	< 0.05	< 0.05	< 0.05	< 0.05	< 0.05	< 0.05	< 0.05	< 0.05
SegDINO	**0.85**	< 0.05	< 0.05	< 0.05	< 0.05	**0.14**	< 0.05	< 0.05

5. 视觉比较

（1）ACDC、LA2018 和 Synapse 数据集上的少样本结果可视化

图 1 展示了在单样本训练设置下，ACDC、LA2018 和 Synapse 数据集在磁共振（第一、二行）和计算机断层扫描（第三行）模态上的典型分割结果。第一列显示叠加在原始图像上的真实标注掩膜，其余各列则依次展示 DINO-AugSeg、nn-UNet、SwinUNETR、SimSiam 和 SegDINO 生成的预测结果。

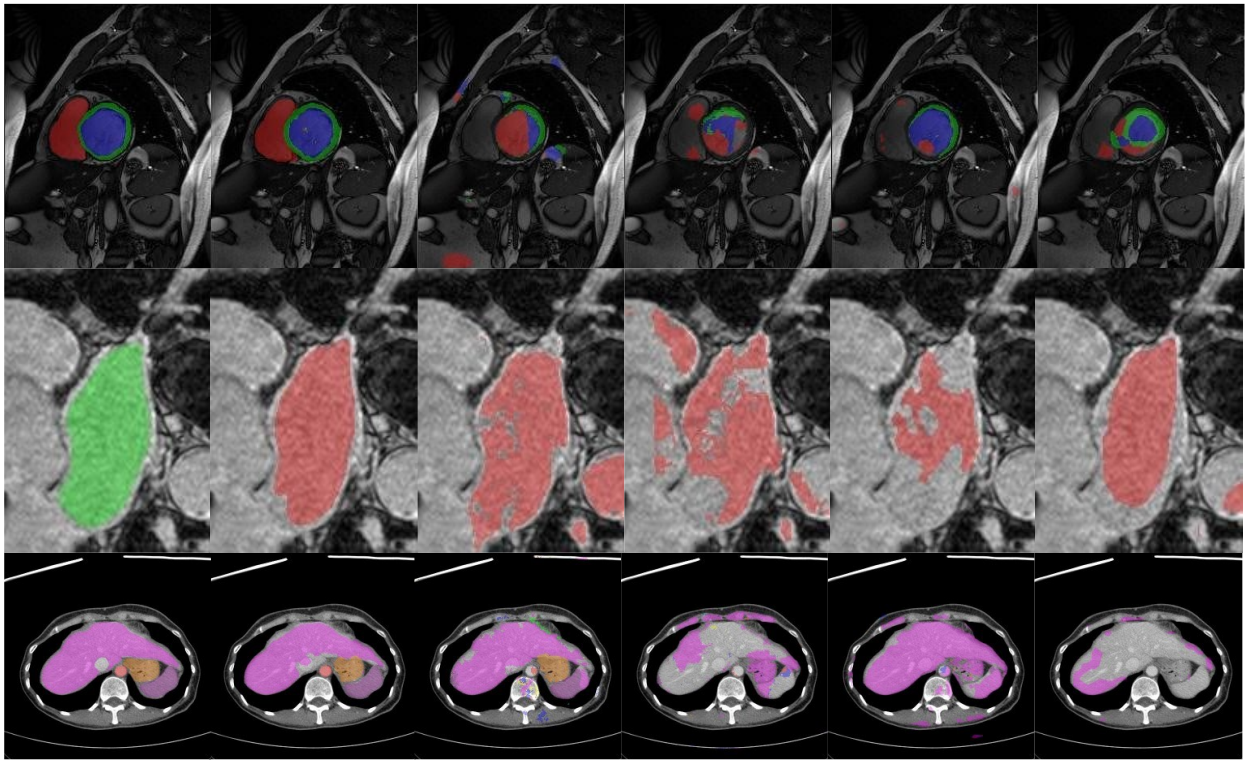


图 1。 展示了 ACDC、LA2018 和 Synapse 数据集在磁共振（第一、二行）和计算机断层扫描（第三行）模态下的代表性分割结果。第一列显示真值叠加图，其余各列展示单样本训练设置下的预测结果。

（2）TN3K、Kvasir-SEG 和 ISIC2018 数据集上的少样本结果可视化

图2 展示了在 TN3K、Kvasir-SEG 和 ISIC2018 数据集（从上到下）上，100样本、40样本和25样本训练设置下的定性比较，重点呈现了所提出的 DINO-AugSeg 的典型分割结果。第一列显示叠加在对应图像上的真实标注掩膜（绿色），后续各列则展示 DINO-AugSeg、nn-UNet、SwinUNETR、SimSiam 和 SegDINO 生成的预测掩膜（蓝色）。在所有数据集和少样本设置下，DINO-AugSeg 均始终表现出更优的分割性能。

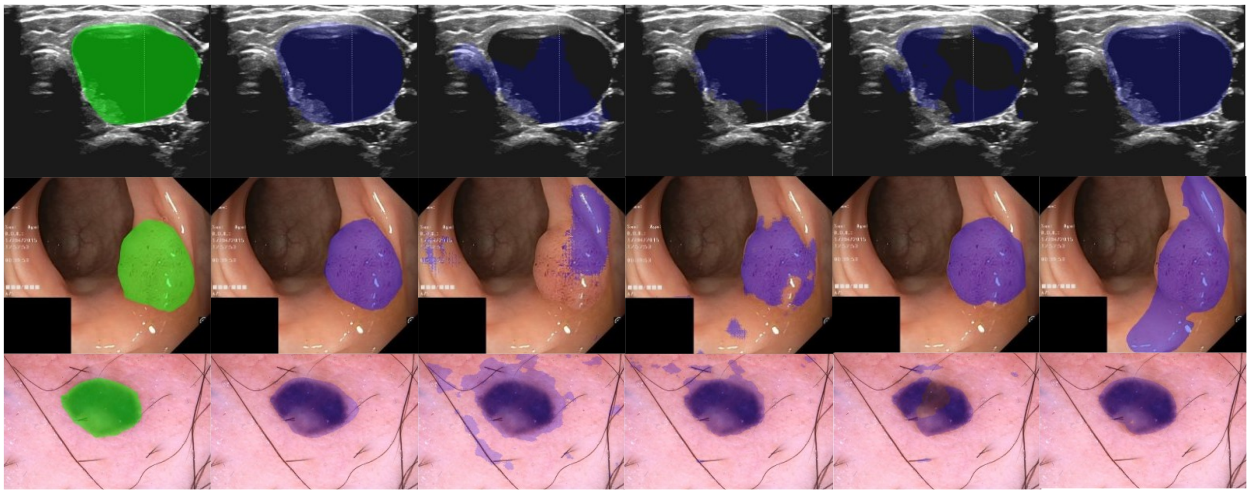


图2。在超声（超声）、内窥镜检查和皮肤镜检查模态下，TN3K、Kvasir-SEG 和 ISIC2018 数据集在 100样本、40样本和 25样本训练设置中的代表性分割结果。真实标注掩膜以绿色显示，模型预测以蓝色显示。