基于掩码自监督图像重建的超声影像基础模型预训练

摘要

超声是一种重要的无创影像学检查，超声智能化是在医学场景中发挥人工智能优势的一个重要应用。近期，围绕基础模型话题大量文章提出了有效的自监督预训练策略，然而在超声成像领域却进展缓慢。本文基于以近期发表在 MedIA 上的一篇超声基础模型研究性论文为启发，首先复现了其中的关键预训练代码并在此基础上提出了一种在图像自回归重建过程中的多任务压缩感知损失的改进策略，在原有方法基础上显著改进了其表征学习方面的性能。通过上游图像预训练与下游多任务测试，我们在图像分割与开源图像分类任务中均取得了有竞争力的指标。这有望给基于掩膜预训练的超声影像预训练方法提供在信息压缩感知理论方面的有效参考并为该领域研究提出新的解决方案。

关键词：超声影像分析；自监督学习；表征学习；模型预训练

1 引言

基础模型旨在以通用数据训练的视角提供对特定数据模态的特征基础表示能力。在通往通用人工智能前进的道路上，基础模型是不可绕开亟待解决的关键问题。近年来，随着自然语言理解研究领域中如 Chat-GPT 这样的大模型的兴起，在图像中也相继出现了具有领域适应性的通用基础模型工作。图像基础模型研究一般遵循将非同质化的海量数据与具有通用特征提取能力的基础范式建立鲁棒联系的研究思路。以自回归、自监督的方式训练模型在海量数据提供的丰富特征空间中挖掘符合当前模态的基础表征从而以低成本、高效能的方式迁移用于多种下游任务。在自然图像领域，最早以讨论自监督学习模型为启发探寻有益于脱离标签下的数据挖掘范式。2021 年 He 等人将自然语言处理中基于掩膜的自回归编码预训练首次用于图像分析并提出了一种基于掩码的自监督学习方法进行模型预训练。他们成功验证了通过添加掩膜而驱动解码器进行图像信息自回归重建的预训练方法在图像领域的适用性，而这样的研究思路也不断迁移到多个学科领域。

超声医学图像是通过收集回波信息进行快速低成本成像的重要临床影像学检查之一。在此之前，超声图像模型多呈现小而粗、杂且繁的特点，对待不同的超声图像识别任务，研究者分别在特定范围的数据集与固定验证方式下设计了许多简单领域适用性的方法。然而，随着临床医学对人工智能模型泛化性的需求，特定数据适用性的小模型往往需要重新训练亦或是增量学习才能获得模型性能的提高。在超声成像的深度学习研究中，尚未提出一种具有模态特征表示能力的基础模型。本文的研究工作正是围绕此而展开。

2 相关工作

受大规模语言模型革命性影响的启发，近期研究广泛聚焦于大规模视觉基础模型，以探索其在通用视觉任务中的潜力 [6]。这些视觉基础模型旨在作为多种视觉任务的通用骨干网络，为理解和处理视觉数据提供坚实的基础。这类模型的通用性源于其在大规模、多样化的数据集上的预训练，这些数据集涵盖了广泛的视觉内容。根据预训练方法的不同，视觉基础模型可分为两类：任务特定（Task-Specific）基础模型和任务无关（Task-Agnostic）基础模型 [1]。任务特定基础模型在大规模标注数据集上进行预训练，以实现对特定任务的广泛适用性。其中最具代表性的工作之一是由 SA-1B 数据集（包含十亿条分割标注）开发的 Segment Anything Model (SAM) [10]。这类模型可以通过简单的提示（Prompting）或微调（Fine-tuning）应用于下游任务。然而，对大规模标注数据的依赖限制了其发展，特别是在标注成本高昂的任务中可行性较低。鉴于未标注数据的广泛存在，任务无关基础模型采用自监督预训练范式，通过学习更复杂的视觉模式和更通用的特征表示，从更大规模的视觉数据库中提取信息。这些自监督预训练主要包括掩码图像建模（Masked Image Modeling, MIM）和对比学习（Contrastive Learning）。其中，MIM 的代表性工作包括 MAE [7] 和 BEiT [3]，对比学习的代表性工作则包括 SimCLR [5] 和 MOCO [8]。这些视觉基础模型在多种视觉任务中展现出显著的灵活性和高效性，尤其在资源受限的场景下表现尤为突出。这些研究的成功为进一步探索和开发适用于不同成像技术的先进视觉基础模型奠定了基础，有望推动视觉数据分析领域的发展，并拓宽其在多种应用场景下的应用。

尽管视觉基础模型在自然图像领域取得了显著成就，并受到广泛关注，但在医学影像领域的研究仍面临诸多挑战 [2]。一方面，不同成像原理导致自然图像领域的成熟方法和基础模型难以直接迁移至医学影像领域。另一方面，医学图像分析任务涉及更复杂、隐性的映射关系，这对基础模型的信息提取能力提出了更高要求。基于上述限制，许多医学影像基础模型根据成像模态的特点进行了特定设计，以识别由于成像原理差异导致的灰度分布和特征的显著差异。例如，MIS-FM [11] 是针对计算机断层扫描（CT）图像开发的基础模型，该模型通过在大规模 3D 体积数据上预训练，在头部、颈部、胸部和腹部等多目标分割任务中表现出色。针对视网膜图像，RETFound [13] 采用 MIM 进行预训练，在眼科疾病诊断中展现了极高的标签效率。而对于内窥镜视频，Endo-FM [12] 通过对比学习建立，分别在胃肠道疾病的分类、分割和检测任务中进行了实验验证。这些研究充分证明了基础模型在各自下游任务中的显著有效性。

3 超声掩膜自编码表征学习

本文主要依赖于 Jiao 等人 [9] 的 USFM 论文对其方法学部分的细节信息进行复现。3.1 章节概述了基于掩膜的自监督学习对于超声图像分析的整体流程。3.2 章节详细说明了原始论文中对于自监督特征提取范式的思考并提出了一种目标函数改进策略。3.3 章节对下游任务的数据任务情况进行了概述。

3.1 方法概述

图1详细展示了本文的工作流程。对于输入的超声图像样本首先将其划分为空间域图像与频域图像，分别从空间域与频率域添加随机掩膜来破坏原始信息通道，然后通过傅里叶反变换将二者在空间域进行逐像素加和，进而将剩余图像块加载到图像编码器中进行特征提取。图像特征编码器将学习被掩膜和频率覆盖处理后的图像的特征信息。由于基于 Transformer 的图像处理方法输入与输出通道之间的同胚关系，因此，解码器将以相同的初始尺寸结合掩膜重新恢复到原始图像。在这个过程中，解码器只能利用掩膜覆盖后的图像空间域与频域的表征而恢复图像。空间域掩膜重建过程遵循 He 等人提出的 MAE 的原始设计，而频谱图由离散傅里叶变换得到与自监督频率标签构成回归性损失。

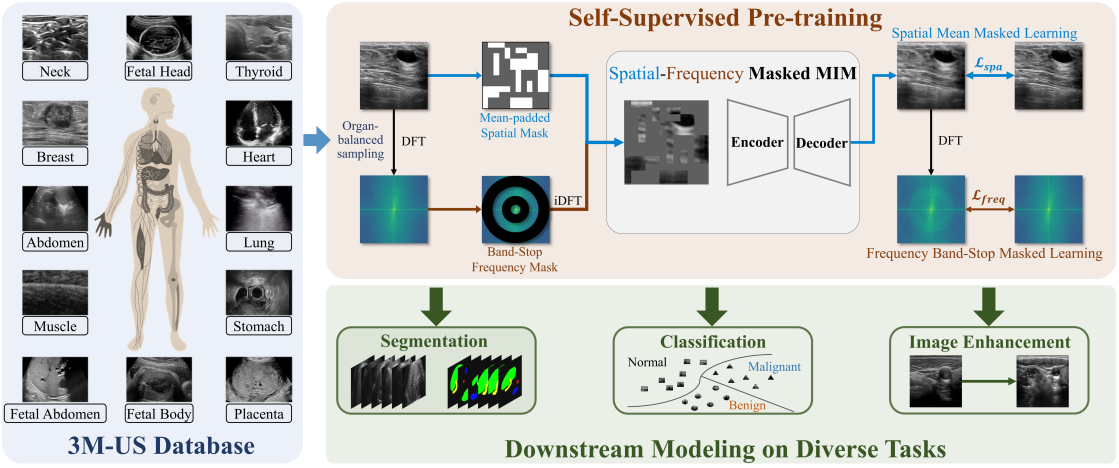

图 1. 复现 USFM 概述

根据 Jiao 等人提出的原始贡献，他们的自监督损失可以定义为：

$$L = L_s + \alpha \cdot L_f \tag{1}$$

其中，L_s 表示空间域损失，采用像素间的绝对误差（Mean Absolute Error, MAE），L_f 表示频率域损失，采用频谱图像素的欧几里得距离（L2 范数），α 为静态超参数。然而，该损失函数存在以下不足：**静态权重问题**：α 是一个固定的超参数，无法根据训练过程动态调整空间域与频域的平衡。**损失形式单一**：空间域使用 MAE，过于关注像素级细节，忽略图像的结构信息；频域采用 L2 范数，对高频噪声过于敏感。**尺度不一致**：L_s 与 L_f 的数值尺度可能存在较大差异，导致优化过程中某一损失主导模型更新。

3.2 改进策略：优化损失形式

针对上述问题，本文提出改进的损失函数，通过引入**结构相似性（SSIM）**损失和**对数频谱（Log Spectral Distance, LSD）**损失，提升模型的特征提取能力和鲁棒性。

在空间域损失改进方面，传统的 MAE 损失忽略了图像的结构信息，难以捕捉纹理与边缘特征。因此，本文采用**结构相似性损失（SSIM）**和 **L1 损失**的加权组合：

$$L_s = \lambda_1 \cdot (1 - \text{SSIM}(I_{rec}, I_{gt})) + \lambda_2 \cdot \|I_{rec} - I_{gt}\|_1 \tag{2}$$

其中：I_{rec} 表示重建图像，I_{gt} 表示原始图像（Ground Truth）。SSIM(·) 用于衡量图像在亮度、对比度和结构上的相似性，更关注图像结构信息。$\|I_{rec} - I_{gt}\|_1$ 为 L1 损失，强调像素级差异，有助于细节恢复。λ_1 和 λ_2 用于平衡 SSIM 和 L1 损失的权重。

在频域方面，原始的频域损失原采用 L2 范数，对高频噪声敏感。为了减小噪声影响，采用**对数频谱损失（LSD）**，在对数域中衡量频谱差异：

$$L_f = \|\log(\mathcal{F}(I_{rec}) + \epsilon) - \log(\mathcal{F}(I_{gt}) + \epsilon)\|_2 \tag{3}$$

其中：$\mathcal{F}(\cdot)$ 表示傅里叶变换，ϵ 是防止对数计算时分母为零的小常数。对数操作能抑制高频噪声，增强对低频和全局结构的关注。将上述改进后的空间域和频域损失结合，得到最终的损失函数：

$$L = \lambda_1 \cdot (1 - \text{SSIM}(I_{rec}, I_{gt})) + \lambda_2 \cdot \|I_{rec} - I_{gt}\|_1 + \lambda_3 \cdot \|\log(\mathcal{F}(I_{rec}) + \epsilon) - \log(\mathcal{F}(I_{gt}) + \epsilon)\|_2 \tag{4}$$

其中，λ_3 控制频域损失在总损失中的权重，λ_1。改进的自监督损失在三个充分发挥了原始设计的自监督特征提取潜力。**结构信息保留**：引入 SSIM 损失增强了模型对图像结构和纹理的建模能力。**鲁棒性增强**：对数频谱损失有效抑制高频噪声影响，使模型更关注全局结构信息。**细节与全局信息平衡**：L1 损失关注局部细节，SSIM 和 LSD 共同提升模型对全局和局部特征的学习能力。

3.3 下游任务测试

自监督学习的重要应用之一在于将预训练过程中学到的特征表示迁移到下游任务中，以提升模型在特定任务上的性能。为了验证本文方法的有效性，我们选择了两个具有代表性的超声图像下游任务进行测试。首先，在 TN3K 分割任务中，模型需对甲状腺超声图像中的结节区域进行精确分割，该任务对模型的特征提取能力和局部细节建模提出了较高要求。其次，我们还在胎儿脑超声数据集（FBUSD）上进行了多分类任务实验，该任务涉及对不同超声切面的精确分类，考验模型对全局结构和局部特征的理解能力。通过将自监督学习获得的特征参数迁移至上述任务，进一步验证了本文提出方法在超声图像分割与分类任务中的泛化能力和优越性能。

4 实施细节

由于 Jiao 等人初始提出的数据集闭源设置，因此我们使用了其公开的权重进行了损失提升策略下的自定义数据集再训练。我们从华南地区某医院收集了约 12 万张包含 15 器官类别的超声多器官数据集 (Ultrasound Muti-task Dataset,UMTD)，将其随机的以 4:1 划分为训练集和测试集，然后使用上述方法进行自监督预训练。

4.1 实验设置

所有实验在 4 张 NVIDIA RTX 4090（24G）完成，使用 torch1.13 框架与 huggingface transformer4.45 外部依赖复现了相关代码。设置批次大小为 256，以 Jiao 等人开源的权重为 ViT-b 的训练起点，使用拟改进的损失函数进行模型微调训练。其中超参数经过多次实验选

表 1. 基于自监督的超声图像预训练上游表征测试结果

方法	特征编码框架	Top1 精度	Epoch
MAE	ViT-B	87.76 ± 0.13	300
USFM	ViT-B	88.35 ± 0.37	300
USFM^{+}(Ours)	ViT-B	89.07 ± 0.52	300

表 2. 在 TN3K 上迁移学习测试下游任务分割结果

特征编码器	方法	IOU↑	Dice↑
ViT-B	Random	47.13	64.06
ViT-B	Pretrained	50.30 (+3.17)	66.94 (+2.88)

择了最优值：λ_1 设置为 0.5，λ_2 为 0.5，λ_3 为 0.1。使用动态的学习率调整策略并进行了 200 个 Epoch 的训练。训练 10 个 epoch 大概需要 25 分钟，100 个 epoch 花费约 4 个小时，训练 200 个 epoch 需要约 8 个小时。

4.2 主要结果

我们在表1展示了包含原始 MAE 在内在超声图像数据的上游自监督表征学习测试结果，我们报告了所有五次测试的均值和标准差。结果表明，通过改进的自回归编码损失函数，基于空间域与频率域解耦的超声图像表征模型预训练效果得到了有效提升。

此外，我们使用所获得的超声图像预训练表征迁移学习给 TN3K 分割任务和 FBUSD 图像分类任务进行下游测试。表2和3展示了详细基于 TransUnet [4] 的图像分割结果和基于 ViT-b 的图像分类结果。

5 总结与展望

本文基于 Jiao 等人在 MIA 上发表的关于超声图像基础模型预训练的工作，从解耦图像空间域与频域掩膜重建的角度复现了原始论文提出的自监督学习框架并从压缩编码感知角度提出了一种改进的自回归目标函数，并将其应用于一个自定义的私有数据集进行自监督预训练。为了合理说明拟提议的目标函数改进策略能够显式的加强图像自回归编码模型，本文使用自监督学习预训练获得的参数迁移学习给两个开源任务，分别从语义分割与图像分类角度验证了所获得表征能够有效用于其他开源任务。本文的实验性结论有效表明了使用多重损失约束的图像自回归编码是一种用于超声图像视觉模型有效预训练策略。

表 3. 在 FBUSD 上迁移学习测试下游任务分类结果

特征编码器	预训练	Top1 精度
ViT-B	Random	74.61
ViT-B	US-Pretrained	75.50 (+0.89)

目前，基于 MAE 的图像自回归预训练视觉模型方法是当前主要的基础模型预训练范式之一。但是本文在实验实施中发现，对于超声图像模态来说，由于其具有高噪声，低对比度等特点，基于自回归掩膜重建的自监督学习策略并不是一种最优的预训练范式。本文认为，基于随机挖取并设置图像掩膜而使用剩余图像块表征的进行原始信息通道重建的方式将会破坏超声前景的固有特征分布。这是因为一些超声图像由于扫描超声探头成像的原因，图像一般成像锥形，无意义背景信息约占据 30% 的像素面积，而一般的自监督掩膜设置比例大都在 70% 以上，使用图像均值掩膜或者像素值为 0 的掩码进行图像重建均无法有效还原原始数据表征。这种基于几何与压缩感知的预训练策略将有损于超声自监督特征学习，挖掘新的范式，比如对比学习或自蒸馏聚类学习仍是该领域有前景的研究方向。

参考文献

[1] Muhammad Awais, Muzammal Naseer, Salman Khan, Rao Muhammad Anwer, Hisham Cholakkal, Mubarak Shah, Ming-Hsuan Yang, and Fahad Shahbaz Khan. Foundational Models Defining a New Era in Vision: A Survey and Outlook, 2023.

[2] Bobby Azad, Reza Azad, Sania Eskandari, Afshin Bozorgpour, Amirhossein Kazerouni, Islem Rekik, and Dorit Merhof. Foundational Models in Medical Imaging: A Comprehensive Survey and Future Vision, 2023.

[3] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. BEiT: BERT Pre-Training of Image Transformers, 2022.

[4] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L Yuille, and Yuyin Zhou. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*, 2021.

[5] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A Simple Framework for Contrastive Learning of Visual Representations, 2020.

[6] Ting Chen, Saurabh Saxena, Lala Li, Tsung-Yi Lin, David J. Fleet, and Geoffrey Hinton. A Unified Sequence Interface for Vision Tasks, 2022.

[7] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked Autoencoders Are Scalable Vision Learners. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15979–15988, 2022.

[8] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum Contrast for Unsupervised Visual Representation Learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9726–9735, 2020.

[9] Jing Jiao, Jin Zhou, Xiaokang Li, Menghua Xia, Yi Huang, Lihong Huang, Na Wang, Xiaofan Zhang, Shichong Zhou, Yuanyuan Wang, et al. Usfm: A universal ultrasound foundation model generalized to tasks and organs towards label efficient image analysis. *Medical Image Analysis*, 96:103202, 2024.

[10] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment Anything, 2023.

[11] Guotai Wang, Jianghao Wu, Xiangde Luo, Xinglong Liu, Kang Li, and Shaoting Zhang. MIS-FM: 3D Medical Image Segmentation using Foundation Models Pretrained on a Large-Scale Unannotated Dataset, 2023.

[12] Zhao Wang, Chang Liu, Shaoting Zhang, and Qi Dou. Foundation Model for Endoscopy Video Analysis via Large-Scale Self-supervised Pre-train. In Hayit Greenspan, Anant Madabhushi, Parvin Mousavi, Septimiu Salcudean, James Duncan, Tanveer Syeda-Mahmood, and Russell Taylor, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*, Lecture Notes in Computer Science, pages 101–111, Cham, 2023. Springer Nature Switzerland.

[13] Yukun Zhou, Mark A. Chia, Siegfried K. Wagner, Murat S. Ayhan, Dominic J. Williamson, Robbert R. Struyven, Timing Liu, Moucheng Xu, Mateo G. Lozano, Peter Woodward-Court, Yuka Kihara, Andre Altmann, Aaron Y. Lee, Eric J. Topol, Alastair K. Denniston, Daniel C. Alexander, and Pearse A. Keane. A foundation model for generalizable disease detection from retinal images. *Nature*, 622(7981):156–163, 2023.