Ultrasound Imaging; Foundation Model; Image Analysis; Intelligent Healthcare

Abstract

This study introduces a Universal Ultrasound Foundation Model (USFM) aimed at enhancing the application and label efficiency of ultrasound image analysis in intelligent healthcare. The USFM undergoes self-supervised pre-training on a large-scale multi-organ database, 3MUS, to address challenges such as insufficient databases, image quality, and feature effectiveness in ultrasound image analysis. It employs a spatial-frequency dual masking method to enhance feature extraction. Experiments demonstrate the USFM's generalization ability and high label efficiency across segmentation, classification, and enhancement tasks. Although its performance in clinical settings has not met expectations, the USFM still lays a foundation for the rapid development of ultrasound models. Future work will focus on improving the model's adaptability and generalization in clinical scenarios.

Keywords: Ultrasound Imaging; Foundation Model; Image Analysis; Intelligent Healthcare.

1 Introduction

Ultrasound (US) imaging, recognized for its non-invasive, safe, and widely accessible nature, is extensively used in medical diagnostics and therapeutic interventions. With the advancement of US imaging technology, the integration of artificial intelligence has driven significant progress in the automatic analysis of US images, particularly in tissue segmentation, tumor detection, disease diagnosis, and treatment planning. However, as US imaging expands to include more organs and diseases, traditional US models face challenges such as low labeling efficiency and poor adaptability. Therefore, there is an urgent need for a universal and efficient model that can quickly adapt to different tasks and organs, facilitating its widespread implementation in clinical practice.

In recent years, foundation vision models (such as masked image modeling and contrastive learning) have achieved remarkable success in the natural image domain, demonstrating the immense potential of self-supervised learning on large-scale unlabeled datasets. However, due to significant differences in imaging principles and feature representations between natural and US images, directly transferring foundation models from natural images to US images poses substantial challenges. As a result, developing a foundation model specifically designed for US images has become a critical area of research.

2 Related works

2.1 Visual foundation model

Inspired by the revolutionary impact of large-scale language models, recent research has extensively focused on large-scale visual foundation models and explored their potential in general vision [4]. These visual foundation models are designed to serve as a universal backbone for various visual tasks, providing a solid foundation for understanding and processing visual data. The universality of these models stems from their pre-training on large-scale, diverse datasets encompassing a broad range of visual content. Based on their pre-training approach, visual foundation models can be categorized into two types: task-specific foundation models and task-agnostic foundation models [2]. The former is pre-trained on large annotated datasets to achieve broad applicability for a special task.One of the most representative works is the segment anything model (SAM) [8] developed on a dataset containing one billion labeled segmentation annotations (SA-1B). These task-specific models can be applied through simple prompting or fine-tuning. However, the need for extensive annotated data often limits their development, making them less feasible for tasks with high labeling costs. Considering the abundance of unannotated data, task-agnostic foundation models are established using self-supervised pre-training paradigms to recognize complex visual patterns and learn universal feature representation from larger-scale visual databases. These self-supervised pre-training paradigms are mainly MIM and contrastive learning, where notable works of the former include MAE [6], [5], and the latter includes SimCLR [3], MOCO [10], respectively. These visual foundation models have shown remarkable flexibility and efficiency in various visual tasks, especially in resource-constrained scenarios.The success of these studies paves the way for further exploration and development of advanced visual foundation models tailored to different imaging techniques, holding the promise of advancing the field of visual data analysis and broadening its applications in various contexts.

2.2 Medical image foundation model

Despite the remarkable achievements in vision foundation models on natural images, research in the medical domain remains challenging [10, 12]. The different imaging principles among various medical modalities make it hard to share established foundation models directly from natural to medical images.Some medical modality-specific foundation models have been presented to handle the different medical image modalities. The computed tomography (CT) foundation model, MIS-FM [16], is pre-training on large-scale 3D volumes, and its efficacy is demonstrated across multiple target segmentations, including head, neck, thoracic, and abdominal. The foundation model RETFound [16] has been built by MIM in retinal images and shows high label efficiency in diagnosing eye diseases. As for endoscopy videos, a foundation model named Endo-FM [13] has been constructed by contrastive learning and experimented on classification, segmentation, and detection of gastrointestinal diseases. Deblurring MAE [7] is pre-training on 10,000 US images of the thyroid by MIM with blurring masking, obtaining the performance improvement on the nodule segmentation. The fundamental goal of these foundation models is to learn universal image features from each medical modality [15]. The universal features facilitate the foundation model robustly adapted to the downstream medical task, with high performance [11, 14] and label efficiency [9]. Given the extensive use of US imaging, further developing a universal foundation model for the US will facilitate the advancement of intelligent US analysis in smart healthcare and broaden its application.

3 Method

3.1 Overview

Figure 1 shows the overall structure of the entire article.This paper proposes the USFM (Ultrasound Foundation Model), which is pre-trained on the world's largest 3MUS ultrasound database. The model employs a novel spatial-frequency dual masking method to effectively tackle issues such as low image quality and organ imbalance. USFM is capable of extracting useful spatial information from low-quality images with noise added through the spatial mask, while the frequency domain mask aids in recovering essential frequency information in US images. This enables the model to learn highly generalizable and effective US features.

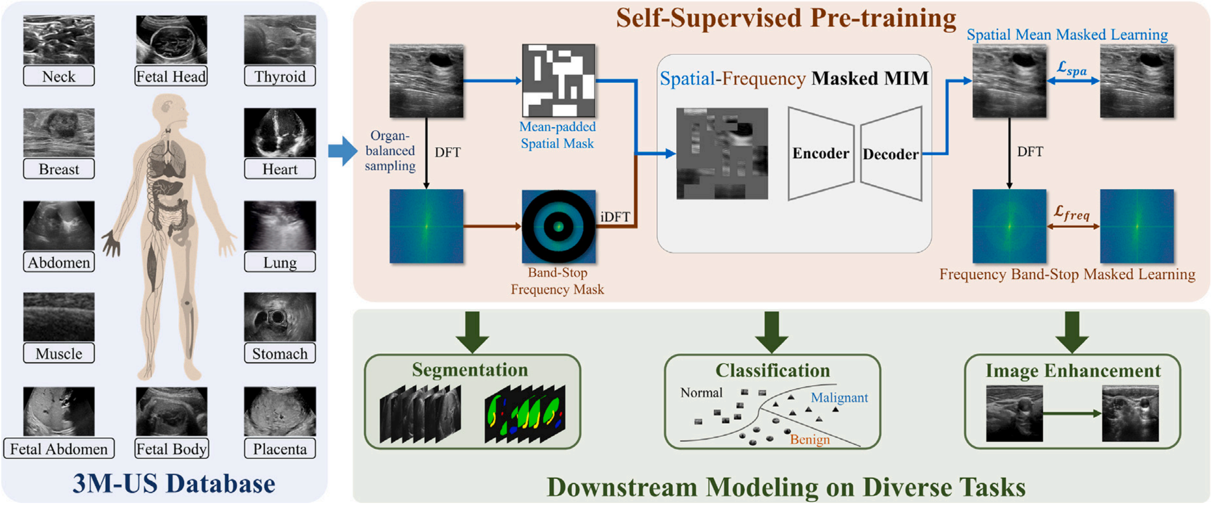

Figure 1. Overview of the method

3.2 Spatial mean masked learning

Unlike existing methods that laboriously eliminate noise, we innovatively continue to add noise to US images in the spatial domain through simple random masking. The USFM is trained in MIM to reconstruct raw images from these masked (noisy-added) images and gain the ability of noise removal and robust feature learning. Specifically, for a US image U, we partition it equally into patches with a given size in the spatial domain. The masked patches are randomly selected based on the masking rate, and their pixel values are replaced by the mean of the image. The spatially averaged masking function can be expressed as

$$M_{\text{spa}}(U) = \begin{cases} \text{mean}(U), & (x, y) \in \text{masked patches} \\ U(x, y), & \text{otherwise} \end{cases} \tag{1}$$

where (x, y) are the coordinates in the spatial domain, and Mean(U) is calculated across the entire image to ensure the continuity of the grayscale distribution. The mask is designed to be mean filling to simulate the noise in US images and avoid the unreasonable disruption of grayscale distribution caused by traditional zero-value (black) filling.

3.3 Frequency band-stop masked learning

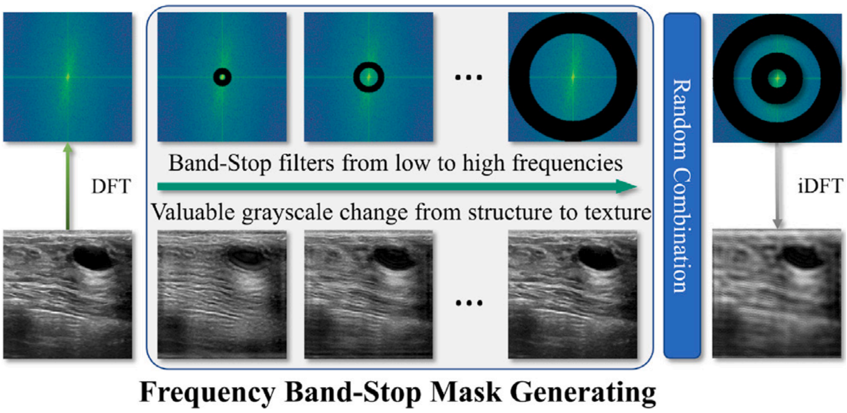
DFT
Band-Stop filters from low to high frequencies
Valuable grayscale change from structure to texture
Random Combination
iDFT
Frequency Band-Stop Mask Generating

Figure 2. Illustrate of frequency band-stop mask generating. The frequency domain distribution is obtained from the spatial domain image by DFT, followed by the generation of multiple band-stop filters ranging from low to high frequencies, corresponding to the valuable grayscale change from structure to texture. These filters are randomly sampled and combined to form the final frequency domain mask.

As shown in Fig. 2, the frequency domain distribution of a US image U of size (H, W) can be obtained through 2D Discrete Fourier Transform (DFT):

$$F(u,v) = \sum_{y=0}^{H-1} \sum_{x=0}^{W-1} U(x,y) \cdot e^{-j2\pi\left(\frac{ux}{H}+\frac{vy}{W}\right)}, \tag{2}$$

where (x, y) represents the coordinates in the spatial domain of US images, and (u, v) denotes the spatial frequency in the frequency spectrum. The $F(u, v)$ represents complex frequency values. According to Euler's formula $e^{j\theta} = \cos\theta + j\sin\theta$, Eq. (2) can also be expressed as:

$$F(u,v) = \sum_{y=0}^{H-1} \sum_{x=0}^{W-1} U(x,y)[\cos 2\pi\left(\frac{ux}{H}+\frac{vy}{W}\right) - j\sin 2\pi\left(\frac{ux}{H}+\frac{vy}{W}\right)]. \tag{3}$$

It is observed that the $F(u, v)$ can be written as $F(u, v) = F_r(u, v) + jF_i(u, v)$ consists of two parts: the real $F_r(u, v)$ and the imaginary $F_i(u, v)$. The amplitude and the phase of the spectrum can be computed from $F(u, v)$ as:

$$|F(u,v)| = \sqrt{F_r(u,v)^2 + F_i(u,v)^2}, \tag{4}$$

$$\angle F(u,v) = arctan\left(\frac{F_i(u,v)}{F_r(u,v)}\right). \tag{5}$$

Within the frequency spectrum of a US image, the amplitude and phase indicate the strength and spatial arrangement of various frequency components, respectively. As seen in the shifted amplitude spectrum on the left side of Fig. 4, the frequency strength gradually goes from low to high from the center outward, and large amplitudes (brighter colors) suggest that these frequency components are vital in US images. The low-frequency components are in the center of the amplitude spectrum, revealing slow-shifted structural information about morphology and deformation, which is essential for US tasks like segmentation and detection.

In contrast, the high-frequency components are in the periphery, detailing the rapidly varied textural information and reflecting developmental progressions and pathological alterations, which are vital for US tasks like maturity measurement and tumor staging.

To capture frequency information, we propose a frequency band-stop masked learning method in the MIM framework of USFM. As illustrated in Fig. 4, the frequency of the US image is randomly masked in the spectrum from low to high frequencies through various band-stop filters. USFM is trained in MIM to extract valuable information across the entire frequency spectrum to recover these masked crucial components. The introduction of frequency band-stop masking enhances the ability of USFM to extract effective US features, which will greatly improve its application in US downstream tasks. The ability of USFM to extract effective US features, which will greatly improve its application in US downstream tasks. For a US image U, the frequency band-stop masking function can be represented as:

$$M_{\text{psd}}(U) = \begin{cases} 0, & f_1 < \sqrt{u^2 + v^2} < f_2, \\ F(u, v), & \text{otherwise} \end{cases} \tag{6}$$

where the f_1 and f_2 are the lower and upper cutoff frequencies, respectively. As illustrated in Fig. 4, the final frequency mask M_{freq} is a combination of several band-stop filters to enrich its diversity.

In summary, the spatial-frequency dual masking in the MIM of USFM applies the following operations to the input image U, forming a dual-masked input U_{dm}:

$$U_{\text{dm}} = \begin{cases} iDFT(M_{\text{freq}}(DFT(U))), & M_{\text{psd}}(U) = 0 \\ M_{\text{psd}}(U), & \text{otherwise} \end{cases} \tag{7}$$

That means the spatial and frequency masking are performed on the input US image, respectively. The spatial masked patch in the frequency masked image will be replaced with the $M_{\text{psd}}(U)$.

3.4 Optimization of self-supervised pre-training

The self-supervised pre-training of our USFM is achieved by the MIM with spatial-frequency dual masking. During pre-training, the U_{em} is generated by the dual masking on organ-balanced sampled US image, and then input to the encoder (E) and decoder (D) structures.

$$U_{rec} = D(E(U_{em})). \tag{8}$$

and the corresponding reconstructed frequency spectrum is:

$$F_{rec}(u, v) = DFT(U_{rec}). \tag{9}$$

The loss of our USFM in the MIM-based self-supervised pre-training phase is

$$\mathcal{L}_{USFM} = \mathcal{L}_{spa} + \lambda \mathcal{L}_{freq}, \tag{10}$$

where λ is the scaling factor to adjust the weight of the two reconstruction losses in the training process. The $\mathcal{L}_{spa}$ and $\mathcal{L}_{freq}$ are the reconstruction loss in the spatial and frequency domain.

In the spatial domain, drawing from prior studies on MIM, we introduce an L1 loss for $\mathcal{L}_{spa}$ to supervise USFM restoring the raw image at the pixel level.

$$\mathcal{L}_{spa} = |U_{rec} - U|. \tag{11}$$

In the frequency domain, the amplitude and phase values of different spatial frequencies exhibit significant variability. Training with the commonly used loss functions, like L1 or L2, which aim to minimize the overall discrepancy, the model tends to align easy spatial frequencies while ignoring hard spatial frequencies. To address this limitation, we employ a focal frequency loss (Jiang et al., 2021) for $\mathcal{L}_{freq}$:

$$\mathcal{L}_{freq} = \frac{1}{HW} \sum_{u=0}^{H-1} \sum_{v=0}^{W-1} w(u,v) \left| F_{rec}(u,v) - F(u,v) \right|^2, \tag{12}$$

where the $F(u,v)$ is the frequency spectrum of the original image. This tailored loss function is specifically designed to improve the attention of difficult frequencies by introducing a spectral frequency weight map $w(u,v)$ that assigns weights to each frequency component.

$$w(u,v) = \left| F_{rec}(u,v) - F(u,v) \right|^a, \tag{13}$$

where a is the scaling factor for flexibility ($a = 1$ in our experiments). The frequency weight map $w(u,v)$ is min-max normalized to the value range of [0,1]. For the frequency component of $F(u,v)$ in the spectrum, $w(u,v) = 1$ means that it receives the maximum attention, and $w(u,v) = 0$ indicates that it is not attended and has no contribution to the loss $\mathcal{L}_{freq}$. The specially designed $\mathcal{L}_{freq}$ ensures a more balanced and accurate reconstruction of all relevant frequency components.

4 Implementation details

4.1 Comparing with the released source codes

- Based on the code provided in the original article, the relevant environment was meticulously deployed, successfully replicating the experimental results presented in the paper.

- Building upon the original dataset, one additional datasets were incorporated, with thorough alignment conducted to validate the model's performance on these new datasets.

- Novel evaluation metrics were introduced, enhancing the comprehensiveness of the assessment framework and ensuring a more rigorous validation of the experimental outcomes.

4.2 Experimental environment setup

This article reproduces the experiment under the Ubuntu platform system, using an NVIDIA GeForce RTX 4090.Create python=3.9, use torch 2.4.1, torchvision 0.19.1,torchaudio 2.4.1.The main data set uses the public breast ultrasound data set to perform ultrasound image segmentation and classification tasks. It is placed in the dataset. The format of the data folder is required to be in Imagenet format.The main running and testing commands, the main details can be viewed in the readme.

4.3 Main contributions

I added a new clinical data set to test the performance of this basic model on downstream classification tasks, and added specificity indicators to improve the evaluation of the model. In practical terms, it can detect doctors' misdiagnosis errors, which is an important indicator. , and then observed that on unbalanced data, the performance is not very good.

5 Results and analysis

The segmentation performance of USFM and ours is reported in Table 1 .The results show that USFM exhibits superior performance across BSUI in US image segmentation tasks.Our results also successfully replicated the effects of usfm. UFSM performed better on BUSI, reaching 0.92 on DSC, 18.05 on HD95, and 0.82 and 0.952 on ACC and SEN.The superior performance demonstrates that USFM has learned critical US knowledge to enhance downstream segmentation tasks by pre-training on a large-scale 3M-US dataset.

Table 1. Comparison results on segmentation tasks.

organs	Model	DSC(%)	HD95	ACC (%)	SEN(%)
BUSI [1]	USFM	84.3 ± 0.17	16.7±20.6	97.34±3.5	83.9 ± 21
	ours	92	18.05	82.0	95.2

The classification performance of USFM and ours is reported in Table 2.Our results on the classification task also successfully replicated the effect of usfm. UFSM performed better on BUSI, reaching 0.878 on ACC, 0.854 on SEN, and 0.907 and 0.897 on PRE and F1.But on other clinical samples, the effect is not very good. On unbalanced data, the f1 score is only 0.48, indicating that although USFM has shown significant potential in the experimental stage, in actual clinical applications, Its performance is not as good as expected.

Table 2. Comparison results on classification tasks.

organs	Model	ACC(%)	SEN(%)	PREC(%)	F1(%)
BUSI [1]	USFM	87.7 ± 1.3	84.4 ± 2.2	89.0 ± 1.1	86.1 ± 1.7
	ours	87.8	85.4	90.7	89.7
Clinical	ours	72.5	40.7	82.9	48.0

6 Conclusion and future work

During the further validation in clinical settings, we conducted an in-depth assessment of the actual performance of the USFM (Universal Ultrasound Foundation Model). Although the USFM demonstrated significant potential during the experimental phase, its effectiveness in real-world clinical applications did not meet the expected standards. This realization suggests that while the USFM's performance in simulated or controlled environments is promising, the complexity of clinical settings, the diversity of patient populations, and the variables in practical operations can all impact the model's performance.To enhance the practical application of the USFM in clinical environments, our future efforts will focus on strengthening the model's adaptability to a variety of clinical scenarios and improving its generalization capabilities by collecting a broader range of datasets. We will also delve into optimization strategies for the model to ensure high accuracy and robustness across different equipment and operator skill levels. Additionally, we will develop interactive learning mechanisms that allow the USFM to adjust in real-time based on clinical feedback and explore the integration of multimodal data to provide more comprehensive diagnostic information.

References

[1] W. Al-Dhabyani, M. Gomaa, H. Khaled, et al. Dataset of breast ultrasound images. *Data in Brief*, 28:104863, 2020.

[2] M. Awais, M. Naseer, S. Khan, et al. Foundational models defining a new era in vision: A survey and outlook. *arXiv preprint arXiv:2307.13721*, 2023.

[3] B. Azad, R. Azad, S. Eskandari, et al. Foundational models in medical imaging: A comprehensive survey and future vision. *arXiv preprint arXiv:2310.18689*, 2023.

[4] H. Bao, L. Dong, S. Piao, et al. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021.

[5] T. Chen, S. Kornblith, M. Norouzi, et al. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, PMLR, pages 1597–1607, 2020.

[6] K. He, X. Chen, S. Xie, et al. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022.

[7] Q. Kang, J. Gao, K. Li, et al. Deblurring masked autoencoder is better recipe for ultrasound image recognition. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 352–362, Cham, 2023. Springer Nature Switzerland.

[8] A. Kirillov, E. Mintun, N. Ravi, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023.

[9] W. Lei, Q. Su, T. Jiang, et al. One-shot weakly-supervised segmentation in 3d medical images. *IEEE Transactions on Medical Imaging*, 2023.

[10] X. Luo, W. Liao, J. Xiao, et al. WORD: A large scale dataset, benchmark and clinical applicable study for abdominal organ segmentation from ct image. *Medical Image Analysis*, 82:102642, 2022.

[11] D. Wang, X. Wang, L. Wang, et al. A real-world dataset and benchmark for foundation model adaptation in medical image classification. *Scientific Data*, 10(1):574, 2023.

[12] G. Wang, J. Wu, X. Luo, et al. Mis-FM: 3d medical image segmentation using foundation models pre-trained on a large-scale unannotated dataset. *arXiv preprint arXiv:2306.16925*, 2023.

[13] Z. Wang, C. Liu, S. Zhang, et al. Foundation model for endoscopy video analysis via large-scale self-supervised pre-train. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 101–111, Cham, 2023. Springer Nature Switzerland.

[14] L. Wu, X. Gao, Z. Hu, et al. Pattern-aware transformer: Hierarchical pattern propagation in sequential medical images. *IEEE Transactions on Medical Imaging*, 2023.

[15] Y. Zhang, J. Gao, Z. Tan, et al. Data-centric foundation models in computational healthcare: A survey. *arXiv preprint arXiv:2401.02458*, 2024.

[16] Y. Zhou, M. A. Chia, S. K. Wagner, et al. A foundation model for generalizable disease detection from retinal images. *Nature*, 622(7981):156–163, 2023.