



USFM: A universal ultrasound foundation model generalized to tasks and organs towards label efficient image analysis

Jing Jiao ^a, Jin Zhou ^b, Xiaokang Li ^a, Menghua Xia ^c, Yi Huang ^a, Lihong Huang ^a, Na Wang ^{a,d}, Xiaofan Zhang ^e, Shichong Zhou ^b, Yuanyuan Wang ^{a,f}, Yi Guo ^{a,f,*}

^a Department of Electronic Engineering, School of Information Science and Technology, Fudan University, Shanghai, China

^b Fudan University Shanghai Cancer Center, Shanghai, China

^c Department of Radiology and Biomedical Imaging, Yale School of Medicine, New Haven, CT, USA

^d SenseTime Research, Shanghai, China

^e Shanghai Artificial Intelligence Laboratory, Shanghai, China

^f Key Laboratory of Medical Imaging Computing and Computer Assisted Intervention of Shanghai, Shanghai, China

ARTICLE INFO

Keywords:

Ultrasound image
Foundation model
Label efficiency
Task adaptability

ABSTRACT

Inadequate generality across different organs and tasks constrains the application of ultrasound (US) image analysis methods in smart healthcare. Building a universal US foundation model holds the potential to address these issues. Nevertheless, the development of such foundation models encounters intrinsic challenges in US analysis, i.e., insufficient databases, low quality, and ineffective features. In this paper, we present a universal US foundation model, named USFM, generalized to diverse tasks and organs towards label efficient US image analysis. First, a large-scale Multi-organ, Multi-center, and Multi-device US database was built, comprehensively containing over two million US images. Organ-balanced sampling was employed for unbiased learning. Then, USFM is self-supervised pre-trained on the sufficient US database. To extract the effective features from low-quality US images, we proposed a spatial-frequency dual masked image modeling method. A productive spatial noise addition-recovery approach was designed to learn meaningful US information robustly, while a novel frequency band-stop masking learning approach was also employed to extract complex, implicit grayscale distribution and textural variations. Extensive experiments were conducted on the various tasks of segmentation, classification, and image enhancement from diverse organs and diseases. Comparisons with representative US image analysis models illustrate the universality and effectiveness of USFM. The label efficiency experiments suggest the USFM obtains robust performance with only 20% annotation, laying the groundwork for the rapid development of US models in clinical practices.

1. Introduction

Ultrasound (US) imaging, recognized for its non-invasive, safe, and widely accessible nature, is instrumental in medical diagnostics and therapeutic interventions (Liu et al., 2019). The widespread accessibility of US imaging has catalyzed its integration with artificial intelligence, driving substantial progress in the automatic analysis of US images, such as tissue segmentation (Xia et al., 2022), tumor detection (Antropova et al., 2017), disease diagnosis (Kang et al., 2022), and treatment planning (Fontanarosa et al., 2015). These successes are attributed to the synergy of data, models and algorithms, namely, fully annotated datasets for specific organs or diseases, specially designed networks and careful training methods. As US imaging expands to new organs and diseases, there is a growing necessity for a US label efficient model generalized to diverse tasks and organs for rapid adaptation and

deployment in medical practice. Addressing this demand will significantly broaden the application of US analysis, facilitating the wider implementation of US models in smart healthcare.

Recently, visual foundation models have emerged as a focal point in the study of natural images, showing potential for creating generalized models (Yuan et al., 2021). These models are often developed on large-scale and unlabeled datasets using self-supervised pre-training methods, such as masked image modeling (MIM) (He et al., 2022) and contrastive learning (Chen et al., 2020b). Benefiting from pre-training on such extensive datasets, these foundation models can derive valuable knowledge from images. The knowledge can be generalized to various downstream tasks and significantly enhance efficiency in model development with outstanding performance, like image classification (Xie et al., 2022), segmentation (Lu et al., 2023), detection (Wei et al., 2022), and generation (Li et al., 2023).

* Corresponding author at: Department of Electronic Engineering, School of Information Science and Technology, Fudan University, Shanghai, China.
E-mail address: guoyi@fudan.edu.cn (Y. Guo).

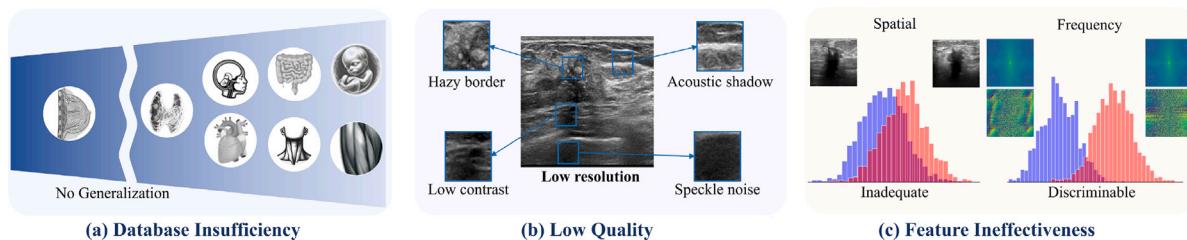


Fig. 1. Three challenges in US foundation model construction. (a) Database insufficiency. (b) Low quality. (c) Feature ineffectiveness.

Inspired by the success in the natural visual domain, introducing foundation models to US images holds promise for the rapid development of models across multiple tasks and anatomical structures. Nonetheless, it is hard to share foundation models from natural to US images due to the inherent differences in imaging principles (Zhang and Metaxas, 2024). Hence, there is an urgent need to develop a foundation model tailored for US images, with comprehensive evaluations of its versatility and adaptability in various downstream tasks.

Due to the unique imaging characteristics and data composition of US images, there are three major challenges that need to be addressed in the construction of the US foundation model, as shown in Fig. 1:

- Database Insufficiency:** The sufficient and diverse US image database is the basis for a comprehensive US foundation model. Gathering US images from multiple sources that cover a wide range of organs using various imaging devices, is difficult due to complex data acquisition and patient privacy concerns. Additionally, the organs with higher morbidity rates are dominant in the database, leading to organ imbalance.
- Low Quality:** The ability to extract meaningful information unaffected by noise ensures the applicability of the foundation model on low-quality US images. US images often have inherent quality issues such as low resolution, contrast, and signal-to-noise ratio (SNR), leading to obscured valuable information. The foundation models built on these noisy and information-sparse US images have difficulty robustly learning US information.
- Feature Ineffectiveness:** Capability of learning effective and general US features is critical for the application of the foundation model on downstream tasks. Extracting clinically relevant and generalizable features from US images is challenging due to the useful US feature is implicit and complex. US features should simultaneously represent the overt spatial features and the underlying practical information concealed within, like the frequency information that depicts the subtle texture changes in the image.

In this study, we have developed a universal US foundation model, namely USFM, with high organ applicability, task adaptability, and label efficiency. To address the above three challenges in US foundation model building, we first established the largest Multi-organ, Multi-center, and Multi-device US database worldwide to date, 3M-US, comprising over two million US images from 12 different human organs. The USFM is pre-trained on the sufficient 3M-US database by employing an organ-balanced sampling strategy to mitigate the organ imbalance. For self-supervised learning on low-quality US images, we introduced a novel MIM-based spatial-frequency dual masking method to learn the generalizable and effective US features. The spatial mask enables the USFM to robustly extract useful spatial information from low-quality US images by recovering original images from the masked (noise-added) inputs. The frequency domain mask is designed to guide the USFM in reconstructing essential frequency information in US images for learning implicit, highly generalizable US representations. The pre-trained USFM can be plug-and-play implemented for various US downstream tasks. Extensive experiments were conducted in our study and have demonstrated its strong generality, superior performance, and excellent label efficiency.

The main contributions of this paper are as follows:

- To the best of our knowledge, we are the first to explore and build a US foundation model, namely USFM, which can be used as a plug-and-play module to improve the efficiency and performance of US image analysis across diverse organs and tasks. Facing the insufficient database issue in the development of USFM, we have constructed the largest and most comprehensive 3M-US database to date, including more than two million US images from diverse organs, centers, and devices.
- To overcome the inherent organ imbalance issue within the 3M-US database, we employed the organ-balanced sampling strategy on the 3M-US database to construct an organ-applicable USFM. Tackling the challenges of low quality and feature ineffectiveness, we introduced a novel MIM-based spatial-frequency dual masking method in the self-supervised pre-training phase of USFM, where spatial masking and frequency masking are proposed to synergistically extract effective US image features, robustly.
- Massive experiments were conducted to comprehensively validate the robust generalizability and superior performance of the USFM across various tasks and organs. Label efficiency experiments demonstrated that USFM maintains practical performance even with minimal annotated data. Additionally, ablation studies confirmed the effectiveness of our proposed organ-balanced sampling and spatial-frequency dual masking method in addressing the challenges of US foundation model building.

2. Related work

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 (Chen et al., 2022). 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 (Awais et al., 2023). 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) (Kirillov et al., 2023) 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 (He et al.,

2022), BEiT (Bao et al., 2022), and the latter includes SimCLR (Chen et al., 2020a), MOCO (He et al., 2020), 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 (Azad et al., 2023; Luo et al., 2022). 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 (Wang et al., 2023b), 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 (Zhou et al., 2023) 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 (Wang et al., 2023c) has been constructed by contrastive learning and experimented on classification, segmentation, and detection of gastrointestinal diseases. Deblurring MAE (Kang et al., 2023) 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 (Zhang et al., 2024). The universal features facilitate the foundation model robustly adapted to the downstream medical task, with high performance (Wang et al., 2023a; Wu et al., 2024) and label efficiency (Lei et al., 2024). 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. Materials and methods

Fig. 2 is the overview of our USFM framework. First, a large-scale and comprehensive 3M-US database containing abundant US images from 12 human organs was constructed. Then, images were sampled and fed into the USFM via an organ-balanced sampling strategy to eliminate potential model bias due to organ imbalance. The pre-training of USFM, based on the MIM framework, integrates a novel spatial-frequency dual masking scheme. This scheme consists of the spatial mean masked learning and the frequency band-stop masked learning. The former simulates common noises in US images, enabling USFM to extract robust and meaningful spatial information. In complementary, the latter facilitates learning of effective implicit features of US images via recovering masked frequency spectrum. Finally, the pre-trained USFM can be conveniently employed for various downstream tasks involving diverse organs and clinical diseases.

3.1. The 3M-US database

A comprehensive large-scale dataset is the prerequisite for developing foundation models with general applicability. In our study, we have established the 3M-US, the largest, Multi-organ, Multi-center, Multi-device US database to date. As illustrated in **Fig. 3**, the 3M-US database includes 2,187,915 US images of 12 common organs involved in human body screening. The database originates US images from various centers and devices worldwide. By encompassing as many US

images as possible, the 3M-US forms the data foundation for achieving organ universality and strong task generalization in our USFM.

All US images included in the 3M-US database underwent the following selection and preprocessing: (1) Images from private databases were used with ethical review approval, and those from public datasets were permitted for public use; (2) Privacy-sensitive information was removed from files and images to ensure data anonymity; (3) All images were converted to grayscale and resized to $(H, W) = (256 \times 256)$ using bilinear interpolation, then saved in portable network graphic (PNG) format with 8-bit depth (i.e., grayscale images) and no compression to preserve the original intensity values. The private dataset was collected from multiple medical centers from 2012 to 2022, including US images of healthy individuals and patients. Most of the images are unlabeled.

The application of US imaging differs depending on the specific organ, resulting in a significant organ imbalance within the 3M-US database, as shown in **Fig. 3**. To prevent the model from overlearning the major organs and neglecting the minorities, we employed a weighted organ-balanced resampling strategy. Specifically, a sampling weight w_{sample} corresponding to different organs was maintained during pre-training on the 3M-US database:

$$w_{sample} = [\frac{1}{\sqrt{N_{organ1}}}, \frac{1}{\sqrt{N_{organ2}}}, \dots, \frac{1}{\sqrt{N_{organ12}}}], \quad (1)$$

where N_{organ} represents the number of images for each organ in the 3M-US database. During each input fetch, the image is sampled from the various organs using weights w_{sample} . As a result, images of minority organs will be resampled more frequently to prevent model bias. Moreover, we performed image augmentation on the input data to increase the diversity of the resampled image for model generalization. The augmentation methods include random rotation, scaling, cropping, brightness and contrast adjustment, and Gaussian blur.

3.2. Spatial-frequency dual masked image modeling

In the pre-training phase, we adopt the MIM-based self-supervised pre-training approach to fully utilize the large amount of unannotated data in the 3M-US database, as illustrated in **Fig. 2**. The key to functional MIM-based self-supervised pre-training is the optimal design of the masking strategy tailored to the characteristics of the US images. As we have summarized, the characteristics that need to be considered are the low quality and difficulty in effective feature extraction. To address these issues, we propose a spatial-frequency dual masked MIM in the pre-training of USFM, consisting of the spatial mean masked learning and the frequency band-stop masked learning, as illustrated in **Fig. 4**.

3.2.1. Spatial mean masked learning

Noise in low-quality images, such as acoustic shadows and artifacts, limits robust learning for meaningful US features. To mitigate the impact of noise, we propose spatial mean masked learning in USFM. 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:

$$M_{spa}(U) = \begin{cases} mean(U), & (x, y) \in \text{masked patches} \\ U(x, y), & \text{otherwise} \end{cases} \quad (2)$$

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.

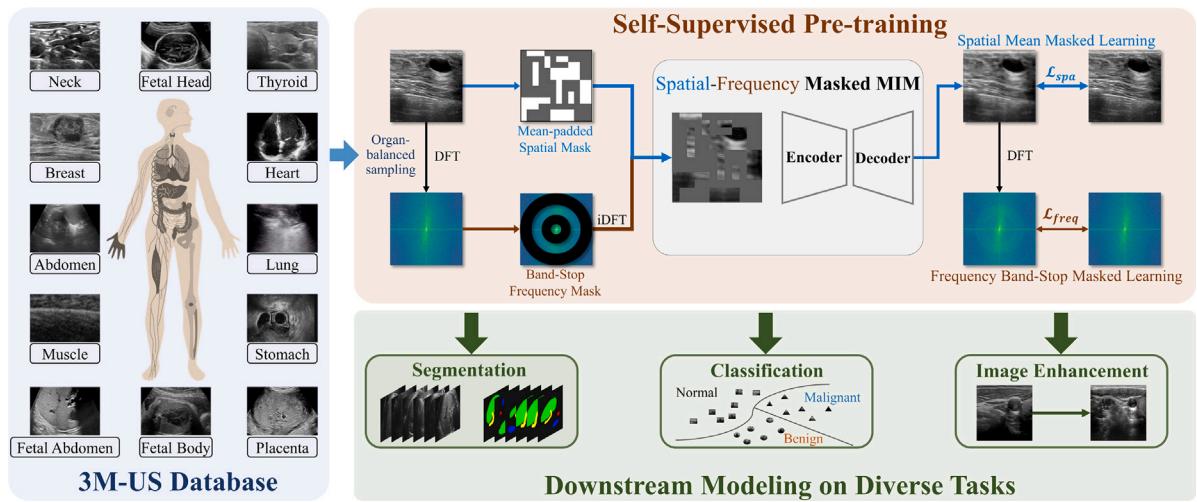


Fig. 2. Overview of the proposed USFM.

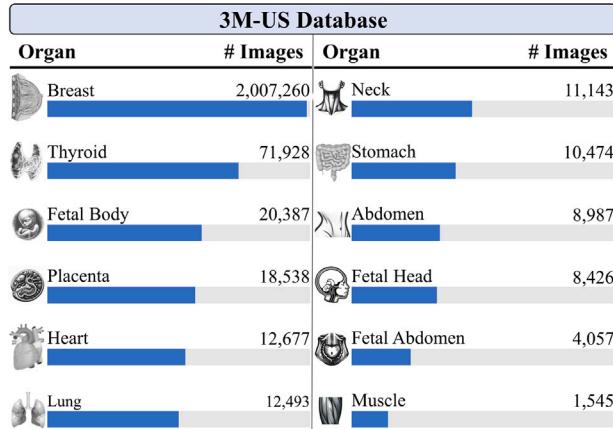


Fig. 3. Summary of the 3M-US database and distribution of involved organs.

3.2.2. Frequency band-stop masked learning

The extraction of effective US features is fundamental for the high performance of USFM in downstream tasks, while it is often challenging in the spatial domain. US image information in spatial is commonly scarce due to low resolution, low SNR, and sparse information density (limited dynamic range of pixel values). Complementary to the spatial domain, the frequency domain of US images contains richer information, where each component is computed based on global information, reflecting both high-frequency texture variations and low-frequency tissue deformations.

As shown in Fig. 4, the frequency domain distribution of a US image U of size (H, W) can be obtained through 2D Discrete Fourier Transform (DFT):

$$F(u, v) = \sum_{x=0}^{H-1} \sum_{y=0}^{W-1} U(x, y) \cdot e^{-i2\pi(\frac{ux}{H} + \frac{vy}{W})}, \quad (3)$$

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 $e^{i\theta} = \cos \theta + i \sin \theta$, Eq. (3) can also be expressed as:

$$F(u, v) = \sum_{x=0}^{H-1} \sum_{y=0}^{W-1} U(x, y) [\cos 2\pi (\frac{ux}{H} + \frac{vy}{W}) - i \sin 2\pi (\frac{ux}{H} + \frac{vy}{W})]. \quad (4)$$

It is observed that the $F(u, v)$ can be written as $F(u, v) = F_r(u, v) + iF_i(u, v)$ consists of two parts: the real $F_r(u, v)$ and the imaginary $F_i(u, v)$.

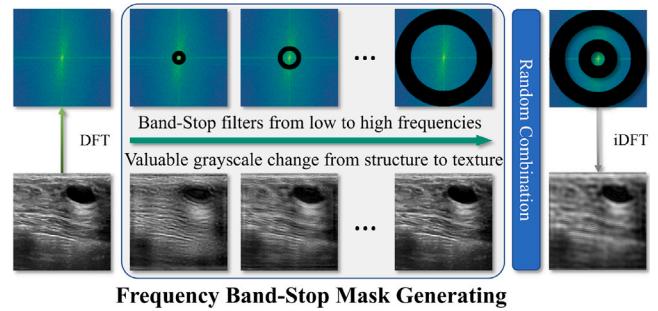


Fig. 4. Illustrate of frequency band-stop mask generating. The frequency domain distribution is obtained from the spatial domain image by DFT, followed by the generation of multiple band-stop filters ranging from low to high frequencies, corresponding to the valuable grayscale change from structure to texture. These filters are randomly sampled and combined to form the final frequency domain mask.

The amplitude and the phase of the spectrum can be computed from $F(u, v)$ as:

$$|F(u, v)| = \sqrt{F_r(u, v)^2 + F_i(u, v)^2}, \quad (5)$$

$$\angle F(u, v) = \arctan(\frac{F_i(u, v)}{F_r(u, v)}). \quad (6)$$

Within the frequency spectrum of a US image, the amplitude and phase indicate the strength and spatial arrangement of various frequency components, respectively. As 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. For a US image U , the frequency band-stop masking function can be represented as:

$$M_{spa}(U) = \begin{cases} 0, & f_1 < \sqrt{u^2 + v^2} < f_2, \dots \\ F(u, v), & \text{otherwise} \end{cases} \quad (7)$$

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} :

$$U_{dm} = \begin{cases} iDFT(M_{freq}(DFT(U))), & M_{spa}(U) = 0 \\ M_{spa}(U), & \text{otherwise} \end{cases} \quad (8)$$

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

3.2.3. 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_{dm} is generated by the dual masking on organ-balanced sampled US image, and then input to the encoder (E) and decoder (D) structures.

$$U_{rec} = D(E(U_{dm})), \quad (9)$$

and the corresponding reconstructed frequency spectrum is:

$$F_{rec}(u, v) = DFT(U_{rec}). \quad (10)$$

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

$$\mathcal{L}_{USFM} = \mathcal{L}_{spa} + \lambda \mathcal{L}_{freq}, \quad (11)$$

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.

$$\mathcal{L}_{spa} = |U_{rec} - U|. \quad (12)$$

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} .

$$\mathcal{L}_{freq} = \frac{1}{HW} \sum_{u=0}^{H-1} \sum_{v=0}^{W-1} w(u, v) |F_{rec}(u, v) - F(u, v)|^2, \quad (13)$$

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.

$$w(u, v) = |F_{rec}(u, v) - F(u, v)|^\alpha, \quad (14)$$

where α is the scaling factor for flexibility ($\alpha = 1$ 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.

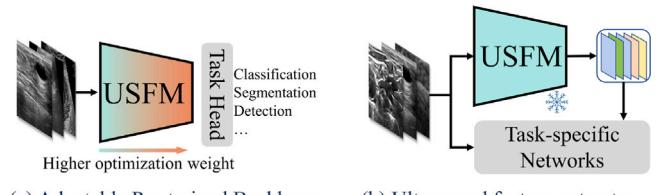


Fig. 5. The diagram of the USFM as a plug-and-play module applied to different tasks: (a) USFM serves as a highly adaptable pre-trained backbone network, which is jointly trained with various task heads. (b) USFM acts as an Ultrasound feature extractor to support task-specific networks.

3.3. USFM adaptation in downstream tasks

Our USFM can be conveniently used as a plug-and-play module combined with existing methods to achieve better results in downstream tasks. As shown in Fig. 5, USFM can be integrated with existing methods in two ways.

Adaptable pre-trained backbone: USFM can act as an adaptable pre-trained backbone combined with different task heads to accomplish various downstream tasks. Leveraging its ability to extract universal features learned from the 3M-US database, USFM facilitates rapid fine-tuning with high performance in downstream tasks. In this scenario, the optimization weights of the downstream model are incremented by layer. The optimization weights for the shallow layers are smaller to maintain the low-level spatial feature extraction capability of the USFM. The deeper layers are optimized with larger weights to extract high-level task-specific features for various downstream tasks, which are different from USFM.

Ultrasound feature extractor: USFM can be integrated into them as a knowledgeable feature extractor for the downstream tasks requiring complex or specifically designed network architectures. In this case, the weights of USFM are frozen and do not participate in network optimization. The performance of the task network will be enhanced by the stable and effective feature extracted in USFM based on the comprehensiveness of the database.

4. Experiments

4.1. Experimental settings

To comprehensively validate the effectiveness and applicability of our developed USFM across downstream task datasets of various organs and diseases, we conducted a series of experiments from four aspects:

4.1.1. Visualization of USFM pre-training

To verify the effectiveness of organ-balanced sampling and spatial-frequency dual masked MIM in pre-training, we presented the masked image and reconstructed results on the spatial and frequency domain of each organ within the 3M-US database. Moreover, we visualized the distribution of different organs in the USFM feature space. First, 100 US images from each of 12 different organs were randomly selected from the 3M-US database, and the features of these images were extracted using USFM. Then, these high-dimensional features were converted to 2-dimensional by UMAP (Ghojogh et al., 2023) for visualization. This experiment can verify whether the features are distinguishable across organs without collapsing.

4.1.2. Comparison on diverse downstream tasks and organs

Experiments for downstream tasks were conducted on multiple organ datasets, covering three common tasks in US image analysis: segmentation, classification, and image enhancement. The details of the downstream task datasets are provided in Table 1.

Table 1
Details of the datasets included in our downstream task.

Downstream	Datasets	Organs	#Images	Targets
Segmentation	Ultrasound Nerve Segmentation (Montoya and Hasnin, 2016)	Neck	5735	Brachial Plexus vs. Background
	TN3K (Gong et al., 2023)	Thyroid	3493	Thyroid nodule vs. Background
	BUSI (Al-Dhabayani et al., 2020)	Breast	780	Breast tumor vs. Background
	Fetal Abdominal Structures Segmentation (Da Correggio et al., 2023)	Fetus	1588	Artery vs. Liver vs. Stomach vs. Vein vs. Background
Classification	BUSI (Al-Dhabayani et al., 2020)	Breast	780	Benign vs. Malignant vs. Normal
	USAnotAI (Kim-Ann, 2020)	Abdomen	366	Bladder vs. Bowel vs. Gallbladder vs. Kidney vs. Liver vs. Spleen
	Fetal Planes (Burgos-Artizzi et al., 2020)	Fetus	12 400	Abdomen vs. Brain vs. Femur vs. Thorax vs. Maternal cervix vs. Other
Image enhancement	USEnhance2023 (Guo et al., 2023)	Carotid, Kidney, Liver, Thyroid	2479	High quality US images

Segmentation: The segmentation experiments were carried out on four datasets of different organs, including the single-object tasks on Neck (brachial plexus), Thyroid (nodule), and Breast (tumor), as well as a multi-object task on Fetus (abdominal structures of artery, liver, stomach, vein, and Background), as shown in Table 1. USFM was generalized to segmentation tasks as an adaptive pre-trained backbone (Fig. 5(a)) by adding a segmentation task head, UperNet (Xiao et al., 2018), as suggested in BEiT (Bao et al., 2022). USFM is used as an adaptable pre-trained backbone to replace the feature pyramid network (FPN) of UperNet for better US feature extraction. The features of the 3rd, 5th, 7th, and 11th blocks in USFM serve as the outputs of the FPN and are fed into the pyramid pooling module (PPM) for the multiscale fusion and segmentation task. We compared the segmentation performance of USFM with common US image segmentation methods, including CNN-based ResUnet (Zhang et al., 2018) and nnUnet (Isensee et al., 2021) and vision transformer (ViT) (Dosovitskiy et al., 2021) based SegFormer (Safa et al., 2023). To validate the effectiveness of US image knowledge learned by USFM, UperNet without pre-trained weights and foundation model (FM)-based SimMIM (Xie et al., 2022) trained on natural images were also included in the comparison. Common metrics such as Dice similarity coefficient (DSC), Hausdorff distance at 95th percentile (HD95), intersection over union (IoU), accuracy (ACC), and sensitivity (SEN) were used to report segmentation performance.

Classification: Classification experiments were conducted on the multi-class tasks of Breast (tumors of benign and malignant, and normal), Abdomen (organs of bladder, bowel, gallbladder, kidney, liver, and spleen), and Fetus (planes of abdomen, brain, femur, thorax, maternal cervix, and other), as shown in Table 1. USFM was generalized to classification tasks by adding a classification task head, i.e., a linear classification layer, following the adaptive pre-trained backbone (Fig. 5(a)). The commonly used models in medical classification, like ResNet50 (He et al., 2016) and DenseNet121 (Que and Lee, 2018), were included in the comparison. Similar to the segmentation experiments, ViT (Dosovitskiy et al., 2021) without pre-trained weights and FM-based SimMIM trained on natural images were also included to demonstrate the role of US image pre-training in USFM. The classification performance was reported using metrics including ACC, recall (Recall), precision (PREC), F1-score (F1), and Matthew's correlation coefficient (MCC).

Image Enhancement: Enhancement refers to generating high-quality images from low-quality US images using generative models (e.g., CycleGAN (Zhu et al., 2017), the most representative). Accurate image enhancement relies on the model learning essential and practical features in US images. In our experiments, USFM was used as a knowledgeable US feature extractor (Fig. 5(b)) to assist CycleGAN. In each of the two generative directions of CycleGAN, USFM extracted image features are combined with CycleGAN encoder features at the channel level to achieve more powerful image encoding capabilities.

The original CycleGAN and the CycleGAN with the natural image feature extractor of SimMIM were included in the comparison. The dataset, USEnhance2023, containing low and high-quality images of the Thyroid, Carotid, Liver, and Kidney, was adopted for the experiment. Three key metrics were utilized to assess the effectiveness of US image enhancement: the structural similarity index measure (SSIM), local normalized cross-correlation (LNCC), and normalized mutual information (NMI). As for the model output (enhanced US image) and target (high-quality images), SSIM compares the visual quality differences by analyzing their luminance, contrast, and structural similarities. LNCC evaluates the alignment and similarity between image patches within local regions. NMI measures the statistical dependency and mutual information between the intensity distributions of two images, normalized to account for the variability in information content across different image pairs. These metrics comprehensively evaluate US image enhancement performance by considering aspects of image quality, alignment accuracy, and information content.

4.1.3. Validation of label efficiency

The label efficiency experiment aimed to confirm that constructing US models with USFM can substantially reduce the need for annotation on the downstream tasks, thus providing a fundament for the low-cost development of US models. The label efficiency experiment was conducted on the segmentation and classification tasks, as they usually have high annotation requirements. In the experiment, we randomly sampled 20%, 40%, 60%, and 80% data from training sets to develop the models on the downstream tasks. The sampling indices on training sets remain the same across different methods for a fair comparison. These models were tested on the entire test set to compare with models developed on the full training set.

4.1.4. Ablation study

The ablation experiment was conducted to validate the role of the proposed organ-balanced sampling and spatial-frequency dual masked MIM modeling in establishing USFM. USFM-woS and USFM-woF represent the USFM conducted without organ-balanced sampling and frequency masking during pre-training, respectively. Their performance on downstream tasks was tested and compared with USFM.

4.2. Implementation details

Data setup: During model training, all US images were input at the size of 224×224 . The data augmentation methods of random rotations, flips, translations, crops, and scaling were implemented on the downstream tasks. The downstream datasets were not included in the 3M-US database to prevent data leakage.

Pre-training: The encoder of the USFM is ViT, and the decoder is a linear layer to reduce the computation during pre-training, as suggested

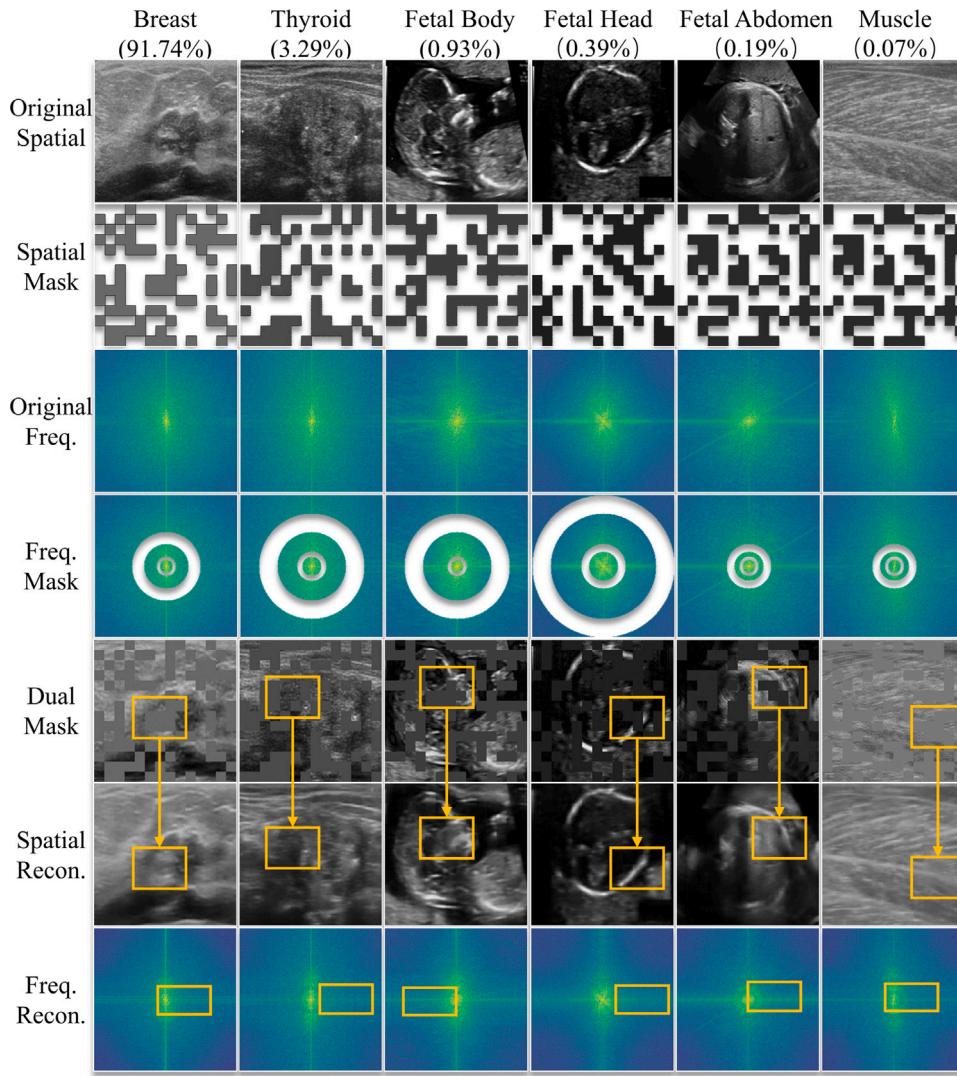


Fig. 6. Examples of the original image, mask, and reconstruction (Recon.) in the spatial and frequency (Freq.) domain of the organs in the 3M-US database. The yellow box highlights the severely masked structure in the US image and the effective reconstruction of the USFM.

by SimMIM. During USFM pre-training, the spatial domain mask ratio is set at 0.4. For frequency masked learning, a region of 10×10 in the center of the spectrum of the US image is first preserved to retain the critical direct current and low-frequency information. Besides the preserved low-frequency region, we set 7 band-stop filters from low to high frequency. The bandwidth of the 7 band-stop filters gradually increases since the information in the high-frequency region is sparser. 2 of these 7 band-stop filters are randomly selected and combined as the frequency mask to improve the masking diversity. The factors λ in Eq. (11) to balance the two reconstruction losses in spatial and frequency domains were set at 0.4 through grid search.

Downstream tasks: The datasets of the downstream tasks are split into training, validation, and test sets as pre-setting at the time of release. The dataset without pre-setting was randomly split in the ratio of 70%, 15%, and 15%. On the segmentation, classification, and image enhancement tasks, all the CNN-based, ViT-based, and FM-based methods are established on the training set and validated on the validation set. The model demonstrating optimal validation performance is saved and evaluated on the test set. The test results are reported as the mean and standard deviation (mean \pm std).

All ViT-based models are implemented by the ViT-b for a balance between performance and efficiency. All models were trained for 200 epochs using the Adam optimizer with a cosine learning rate adjustment strategy. All experiments were conducted on an AMD EPYC 7763

CPU and NVIDIA® GeForce Tesla A100. All models were developed using PyTorch.

5. Results

5.1. Effectiveness of USFM pre-training

Fig. 6 illustrates the spatial and frequency mask and reconstruction results of our MIM-based USFM in the 3M-US database. The USFM exhibits outstanding reconstruction outcomes in both spatial and frequency domains, highlighting its ability in capturing the fundamental and valuable US features necessary in original US images. This adeptness in US feature learning and extraction provides a robust foundation for the function in downstream tasks of USFM.

Thanks to the organ-balanced sampling strategy, USFM maintains a consistent recovery ability for both the majority organs of Breast (91.74%) and Thyroid (3.29%), as well as the minority of Fetal Body (0.93%), Fetal Head (0.39%), Fetal Abdomen (0.19%), and Muscle (0.07%). The capability of organ-universal features extraction further broadens the applicability of USFM in diverse organ contexts.

The spatial mean mask allowed USFM to counteract the inherent noise in US images by reconstructing the original image from artificially introduced noise. As depicted in the yellow box in **Fig. 6**, USFM can successfully recover from noise even in images where critical structures are

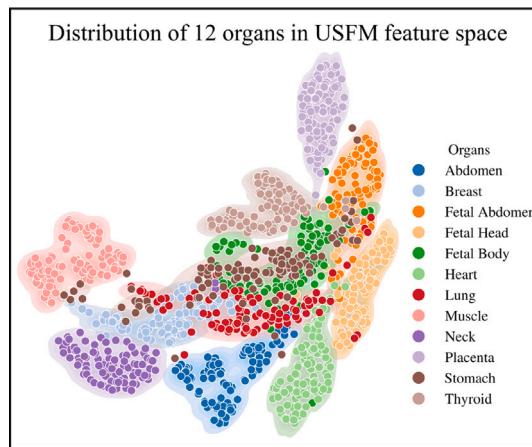


Fig. 7. Visualization of the UMAP distribution of the 12 organs in the USFM feature space.

obscured by noise, such as the masked tumors in Breast, the grayscale distribution in Thyroid and Muscles, and borders in Fetal Body, Head, and Abdomen. This demonstrates the robust feature extraction ability of USFM in low-quality US images.

Frequency masks enable the USFM to extract more effective US features. USFM can recover most of the frequency information in the image even on the aggressive bands-top masking strategy, as shown in Fig. 6 (Freq. Mask vs. Freq Recon.). This indicates that USFM can effectively capture useful frequency information in the image. The US frequency features compensate for the limitation of scarce spatial information, which is especially helpful for downstream tasks requiring a high-level understanding of US images and a detailed focus on texture.

Fig. 7 visualizes the distribution of the 12 organs from the 3M-US database in the USFM feature space. The samples of each organ are clustered in the feature space of USFM. The clustering suggests that USFM can capture the differences in structural and grayscale information from various organs, which is effective for the US downstream. Furthermore, fetal organs, including Fetal Abdomen, Fetal Head, and Fetal Body are close together on the right side of the distribution. Within each organ, the feature representations of different US images are not overlapping, implying that our spatial-frequency dual masked MIM method can prevent USFM from collapsing, i.e., it learns meaningful features rather than noise, which is no definite difference between the images.

5.2. Downstream tasks adaption

5.2.1. Segmentation on various organs

The segmentation performance of USFM for four datasets of different organs is reported in Table 2. The results show that USFM exhibits superior performance across all metrics in US image segmentation tasks. The CNN-based and ViT-based models are inferior to USFM in the segmentation tasks of all organs. Even by increasing the network architecture, the ViT-based SegFormer and UpNet did not show better performance than the CNN-based methods, with the DSCs of SegFormer and UpNet inferior to nnUNet on Neck (79.1%, 76.6% vs. 79.0%), Thyroid (78.9%, 71.1% vs. 80.0%), Breast (65.3%, 66.0% vs. 78.1%), and Fetus (74.2%, 77.9% vs. 82.4%). This result suggests that it is difficult to adequately train large network architectures on small medical US datasets. The gains obtained by increasing the network architecture are smaller than the finetuning of the network.

Compared to the no pre-training methods (CNN-based and ViT-based), FM-based USFM and SimMIM perform better in all organs. Notably, USFM achieves significant improvements over UpNet (the same network architecture but without USFM weight), with DSCs increasing from 76.6% to 80.6%, 71.1% to 85.9%, 77.9% to 85.8%,

and 66.0% to 84.3% on the Neck, Thyroid, Breast, and Fetus, respectively. 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. Due to the fundamental differences between US and natural images, SimMIM, pre-trained on natural images, also shows improved performance in US image segmentation but remains inferior to the USFM. This comparison demonstrates the necessity of the specifically designed USFM for US images in the US segmentation tasks.

Moreover, as seen in Table 2, except for USFM, ResNet, SegFormer, and UpNet, show considerable performance variations across different organs. Among these methods, SegFormer performed best on the Neck (79.1% DSC) and Thyroid (78.9% DSC), but failed on the Breast (65.3% DSC). ResNet only performed best on the Breast (78.5% DSC) and Fetus (81.3% DSC), while its performance on Neck and Thyroid was inferior to SegFormer (78.7%, 77.9%, vs. 79.1%, 78.9% on DSC). As a result, it is difficult to apply any of them as a segmentation method applied generalized to the various organs and tasks. In contrast, our USFM consistently demonstrates superior segmentation performance across all organs and can be used as a universal US image segmentation method.

The segmentation results of the comparison methods are illustrated in Fig. 8. It is evident that USFM outperforms other methods in recognizing the hazy tissue borders and is more robust to noise in the US image. In the segmentation of Neck brachial plexus structure and Breast tumor, ResNet, nnUNet, SegFormer, and UpNet under-segment or over-segment due to their limited ability to resolve blurred edges. Moreover, affected by noise in low-quality US images, UpNet and SimMIM erroneously segment artifacts in Thyroid nodule segmentation. Compared to methods without pre-training, SimMIM performs better by extracting US structure information shared with natural images. However, its performance is still inferior to USFM due to the lack of adaptation to the distinct characteristics of US image.

5.2.2. Classification of diseases and organs

The classification performance of USFM in downstream tasks for different organs is reported in Table 3. Our USFM outperforms other methods across all metrics. It achieved 87.7%, 93%, and 93.6% ACC on the classification of Breast tumors, Abdomen structures, and Fetus planes, respectively, which is sufficient for clinical practice.

On the classification tasks, the ACC of CNN-based DenseNet121 outperforms ViT on Breast (85.7% vs. 68.8%), Abdomen (91.7% vs. 75.0%), and Fetus (91.6% vs. 79.1%) because its concise network architecture can be more adequately trained. The FM-based SimMIM, pre-trained on natural images, was still inferior to DenseNet121 (ACC of 81.2%, 90.0%, and 83.6% vs. 81.2%, 90.0%, and 83.6% on Breast, Abdomen, and Fetus). This result indicates that SimMIM fails to extract useful high-level semantic features in US image classification tasks due to the essential differences between natural and US images. By learning helpful US features from the large-scale US database, our USFM shows a significant performance improvement.

5.2.3. US image enhancement

The image enhancement performance on the USEhance2023 dataset is detailed in Table 4. Compared to the original CycleGAN, using USFM as a knowledgeable auxiliary branch for US features significantly boosts high-quality enhancement performance. USFM achieved superior SSIM compared to CycleGAN: 74.3% over 72.0% in the Thyroid, 75.1% over 73.9% in Carotid, 76.6% over 75.8% in the Liver, and 75.5% over 75.3% in the Kidney. The superior SSIM indicates that the US image features extracted by USFM can help CycleGAN achieve better enhancement results, preserving the real structure and grayscale information. Moreover, USFM achieves superior performance over CycleGAN in LNCC metrics, where are 91.1%, 89.7%, 90.3%, 90.7% vs. 90.7%, 89.3%, 89.2%, 89.7% in the Thyroid, Carotid, Liver, and Kidney, respectively. These results demonstrate the ability of the

Table 2
Comparison results on downstream segmentation tasks.

Organs	Types	Models	DSC (%)	HD95	IoU (%)	ACC (%)	SEN (%)
Neck (Ultrasound Nerve Segmentation)	CNN-based	ResUnet	78.7 ± 15.9	13.4 ± 12.6	67.1 ± 17.1	98.8 ± 8.1	78.9 ± 18.4
		nnUnet	79.0 ± 17.6	12.3 ± 12.0	67.8 ± 17.9	98.2 ± 8.2	78.9 ± 18.4
	ViT-based	SegFormer	79.1 ± 15.5	13.3 ± 12.6	67.5 ± 16.9	98.8 ± 8.0	80.1 ± 18.2
		UperNet	76.6 ± 16.7	15.3 ± 13.9	64.4 ± 17.8	98.7 ± 8.5	76.9 ± 20.4
	FM-based	SimMIM	79.4 ± 15.5	13.3 ± 12.3	67.9 ± 16.7	98.8 ± 0.9	79.9 ± 17.7
		USFM	80.6 ± 14.7	12.3 ± 11.6	69.5 ± 16.3	98.9 ± 0.9	80.7 ± 16.7
Thyroid (TN3K)	CNN-based	ResUnet	77.9 ± 20.1	29.6 ± 32.2	67.4 ± 22.4	95.2 ± 6.1	82.0 ± 21.5
		nnUnet	80.0 ± 21.9	25.6 ± 33.2	70.8 ± 23.4	95.4 ± 6.4	81.7 ± 23.0
	ViT-based	SegFormer	78.9 ± 19.9	27.0 ± 27.4	68.6 ± 21.8	95.4 ± 6.0	81.8 ± 21.7
		UperNet	71.1 ± 22.1	58.2 ± 37.2	59.1 ± 23.7	93.6 ± 7.1	78.0 ± 22.4
	FM-based	SimMIM	83.1 ± 19.1	21.0 ± 24.2	74.5 ± 21.3	96.3 ± 5.5	84.1 ± 20.0
		USFM	85.9 ± 14.6	18.4 ± 24.0	77.5 ± 17.9	96.8 ± 5.0	86.4 ± 16.8
Breast (BUSI)	CNN-based	ResUnet	78.5 ± 22.9	25.6 ± 31.3	69.1 ± 24.1	96.3 ± 5.1	79.2 ± 24.0
		nnUnet	78.1 ± 29.1	24.8 ± 35.8	70.9 ± 29.1	94.3 ± 5.2	80.2 ± 30.0
	ViT-based	SegFormer	65.3 ± 28.0	41.9 ± 38.3	54.1 ± 27.6	94.6 ± 6.1	63.4 ± 29.4
		UperNet	66.0 ± 28.9	43.7 ± 39.3	55.3 ± 28.2	94.8 ± 5.3	66.1 ± 30.4
	FM-based	SimMIM	80.9 ± 24.0	22.3 ± 31.3	72.9 ± 25.6	96.9 ± 4.6	80.9 ± 25.3
		USFM	84.3 ± 17.8	16.7 ± 20.6	76.0 ± 20.3	97.4 ± 3.5	83.9 ± 21.0
Fetus (Fetal Abdominal Structures Segmentation)	CNN-based	ResUnet	81.3 ± 9.9	19.5 ± 5.6	69.5 ± 12.0	98.2 ± 7.1	82.5 ± 12.7
		nnUnet	82.4 ± 12.5	10.8 ± 7.9	71.3 ± 12.2	96.8 ± 7.2	82.4 ± 13.4
	ViT-based	SegFormer	74.2 ± 14.7	22.0 ± 5.9	60.6 ± 14.9	97.8 ± 10.0	69.0 ± 15.6
		UperNet	77.9 ± 11.5	20.6 ± 5.6	65.0 ± 13.1	97.9 ± 9.0	78.5 ± 13.7
	FM-based	SimMIM	83.7 ± 9.1	18.4 ± 4.7	72.8 ± 10.5	98.5 ± 6.0	82.0 ± 10.9
		USFM	85.8 ± 8.5	17.0 ± 4.4	75.9 ± 10.1	98.7 ± 0.6	85.5 ± 9.6

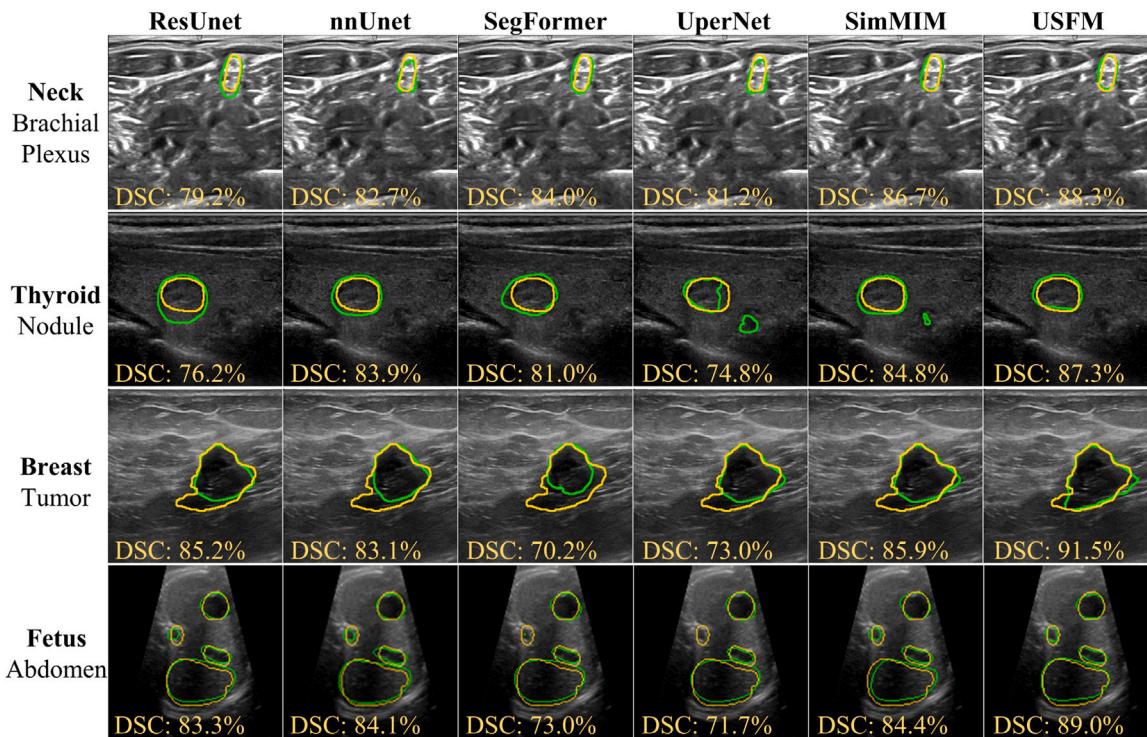


Fig. 8. The segmentation illustration of the comparison methods. Depicted in yellow is the gold truth, and in green is the predicted result.

USFM to learn and preserve critical local information in US images, ensuring the realism of the generated images (higher LNCC with the target image).

The image enhancement example in Fig. 9 demonstrates that the USFM achieves a higher SSIM than CycleGAN and SimMIM. Without the feature auxiliary branch, CycleGAN displays rough and unrealistic textures in Liver and Kidney. SimMIM alleviates the issue of coarse textures while falling in the image contrast (low contrast between organ and background). In contrast to both CycleGAN and SimMIM,

the images enhanced by USFM exhibit finer details and higher contrast, better meeting high-quality US image requirements. It is worth noting that SimMIM using natural image pre-training as an auxiliary branch did not achieve superior performance in SSIM metrics because of its insufficient ability to extract practical information in US images. On the LNCC metric, SimMIM achieved a weak improvement, benefiting from the understanding of the essential information of the image. Compared to SimMIM, our USFM achieved a substantial lead, further illustrating the superiority of USFM for US images.

Table 3
Comparison results on downstream classification tasks.

Organs	Types	Models	ACC (%)	Recall (%)	PREC (%)	F1 (%)	MCC (%)
Breast (BUSI)	CNN-based	Resnet50	77.3 ± 1.7	69.8 ± 2.2	78.1 ± 2.5	72.8 ± 2.2	59.5 ± 3.2
		DenseNet121	85.7 ± 1.6	84.6 ± 2.0	84.8 ± 1.8	84.6 ± 1.8	75.3 ± 2.8
	ViT-based	ViT	68.8 ± 1.0	50.5 ± 1.2	49.3 ± 0.9	48.2 ± 1.1	42.7 ± 2.2
		FM-based	81.2 ± 1.3	79.7 ± 1.4	80.2 ± 1.4	79.9 ± 1.3	67.6 ± 2.2
Abdomen (USAnotAI)	CNN-based	Resnet50	90.0 ± 1.7	90.0 ± 1.7	90.7 ± 1.6	89.8 ± 1.8	88.2 ± 2.0
		DenseNet121	91.7 ± 2.0	91.7 ± 2.0	94.4 ± 1.4	91.6 ± 2.1	90.6 ± 2.1
	ViT-based	ViT	75.0 ± 2.8	75.0 ± 2.8	80.2 ± 2.2	74.1 ± 3.1	71.1 ± 3.2
		FM-based	90.0 ± 1.9	90.0 ± 1.9	91.3 ± 1.6	89.9 ± 1.9	88.3 ± 2.2
Fetus (Fetal Planes)	CNN-based	Resnet50	83.6 ± 0.2	83.3 ± 0.2	79.0 ± 0.2	80.5 ± 0.2	79.7 ± 0.2
		DenseNet121	91.6 ± 0.2	92.6 ± 0.2	89.1 ± 0.2	90.6 ± 0.2	89.5 ± 0.2
	ViT-based	ViT	79.1 ± 0.2	74.9 ± 0.4	74.5 ± 0.3	74.4 ± 0.4	73.5 ± 0.3
		FM-based	89.7 ± 0.2	90.1 ± 0.2	86.8 ± 0.2	88.1 ± 0.2	87.2 ± 0.2
		USFM	93.6 ± 0.2	94.0 ± 0.2	91.5 ± 0.2	92.6 ± 0.2	92.0 ± 0.2

Table 4
Comparison results on image enhancement tasks.

Organs	Models	SSIM (%)	LNCC (%)	NMI (%)
Thyroid	CycleGAN	72.0 ± 3.0	90.7 ± 3.4	21.6 ± 1.7
	SimMIM	71.5 ± 3.0	90.7 ± 3.7	22.3 ± 1.5
	USFM	74.3 ± 2.4	91.1 ± 3.3	23.2 ± 2.1
Carotid	CycleGAN	73.9 ± 2.1	89.3 ± 4.9	21.7 ± 2.0
	SimMIM	73.5 ± 1.9	89.2 ± 5.2	21.2 ± 1.5
	USFM	75.1 ± 2.1	89.7 ± 4.8	23.0 ± 1.8
Liver	CycleGAN	75.8 ± 1.8	89.2 ± 1.7	23.8 ± 1.0
	SimMIM	75.2 ± 3.1	89.5 ± 1.7	23.8 ± 1.3
	USFM	76.6 ± 2.6	90.3 ± 1.6	25.2 ± 1.4
Kidney	CycleGAN	75.3 ± 1.3	89.7 ± 2.3	24.5 ± 1.3
	SimMIM	75.0 ± 2.2	89.7 ± 1.9	24.2 ± 1.3
	USFM	75.5 ± 2.1	90.7 ± 2.0	25.2 ± 1.2

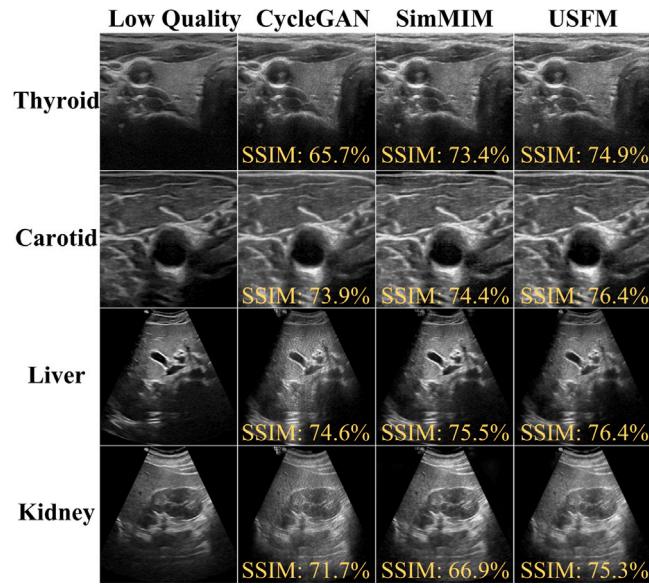


Fig. 9. An image enhancement example of the comparison methods.

5.3. Label efficiency

The label efficiency results for the downstream segmentation and classification tasks are presented in Fig. 10. When developing downstream US on the training set of different label ratios, USFM consistently achieved the highest label efficiency compared to other methods. Even

with only a 20% label ratio available, USFM attain sufficient performance, with DSCs of 72.1%, 76.0%, 72.4%, and 74.2% for the segmentation of brachial plexus in Neck, nodule Thyroid, tumor in Breast, and abdominal structures in Abdomen, as well as F1 of 74.4%, 77.6%, and 81.2% for the classification of tumors in Breast, structures in Abdomen and Plane in Fetus. This result indicates that USFM significantly reduces the need for annotations, facilitating the rapid development of US models.

With the label ratio increasing from 20% to 40%, USFM achieved substantial improvements in almost all organs. For the segmentation tasks on Neck, Thyroid, Breast, and Fetus, the DSCs have increased from 72.1% to 76.4%, 76.0% to 81.4%, 72.4% to 79.7%, and 74.2% to 79.0%, respectively. The F1 of the classification tasks on Abdomen and Fetus have increased from 77.6% to 86.4% and 81.2% to 83.4%. When the label ratio increased to 60%, USFM reached near 100% label ratio performance in the segmentation of Neck, Thyroid, and Breast (Fig. 10(a, b, c)).

In the segmentation tasks, the no pre-trained, ResUnet, SegFormer, and UperNet, showed strong dependence on the amount of annotated data. At a 20% label ratio, these four methods failed to achieve acceptable segmentation performance, with DSCs below 65% in Neck, Thyroid, Breast, and Fetus. Except for USFM, all other methods in classification tasks failed to achieve notable performance at a low label ratio. Particularly at the 20% label ratio, F1 was below 55% in Breast, 75% in Abdomen, and 80% in Fetus. Inadequate performance restricts their application in medical practice, where only partial labels are available. Compared to USFM, SimMIM pre-trained on natural images remained limited in performance due to its weaker feature extraction capability for US images across all label ratios in segmentation tasks. As for the classification tasks, SimMIM performed inferior to the models without pre-training (DenseNet121) at a low label ratio across all organs. This result suggests that SimMIM, pre-trained on natural images, cannot extract high-level US features crucial in US image classification tasks.

5.4. Ablation analysis

Ablation experiments were designed to validate the effects of removing organ-balanced sampling (USFM-woS), and frequency masked learning (USFM-woF) from USFM. In addition, we retrained SimMIM on the 3M-US database, which is called SimMIM-US. SimMIM-US is the baseline to validate the effect of the specially designed pre-training methods for US images in USFM.

The performance comparison of USFM, USFM-woS, USFM-woF, and SimMIM-US in segmentation, classification, and image enhancement tasks is outlined in Fig. 11. Results show a considerable decrease in

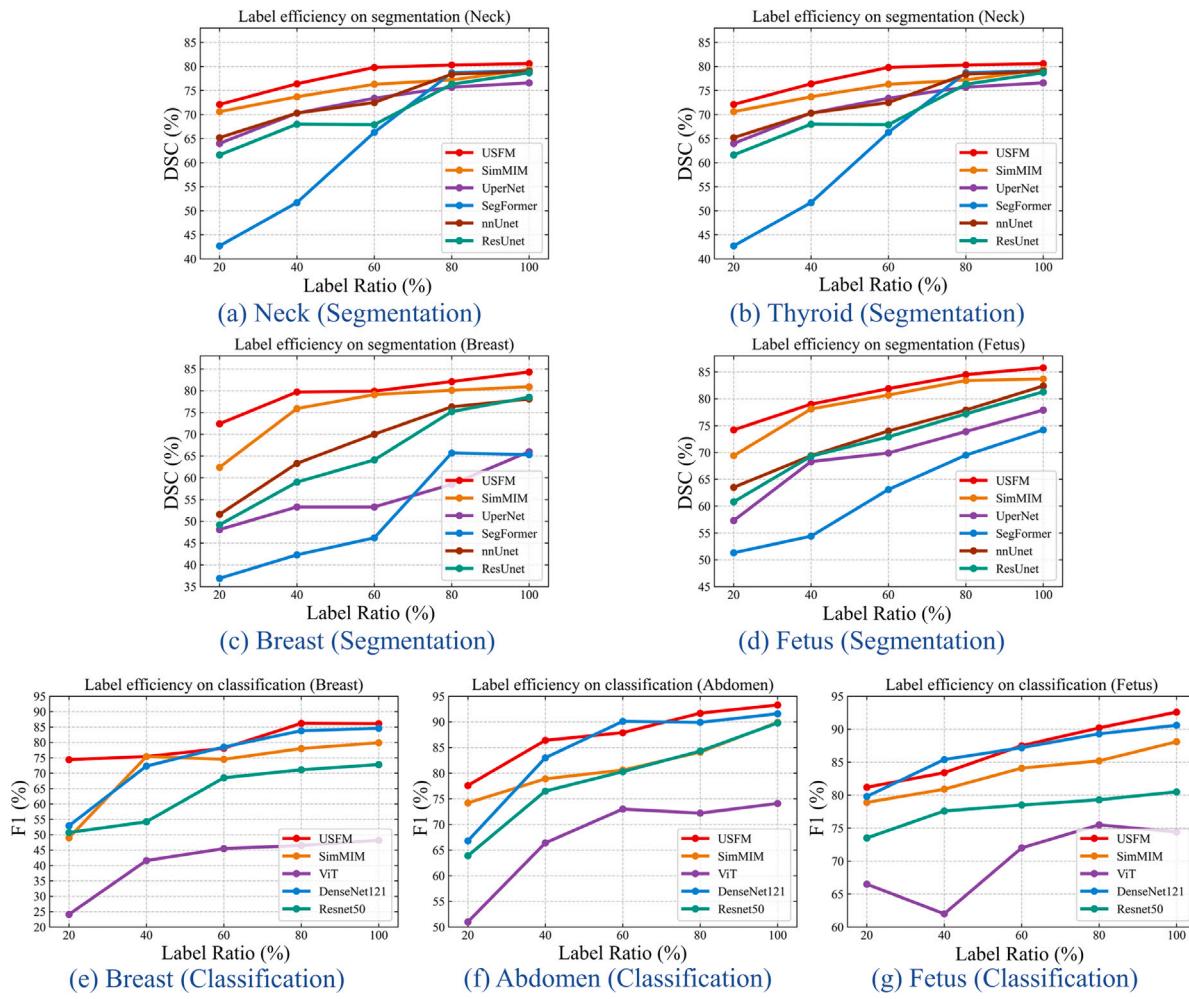


Fig. 10. Label efficiency experiments of the downstream segmentation and classification tasks. The mean DSC (in segmentation) and mean F1 (in classification) of the model trained at different label ratios are reported.

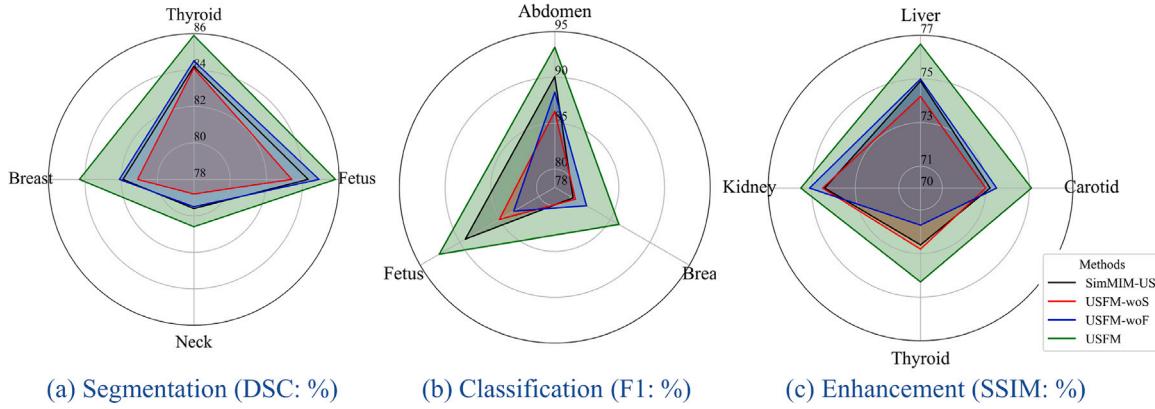


Fig. 11. Ablation study on the segmentation, classification, and detection tasks in all organs. The mean DSCs on the segmentation task, mean F1 on the classification tasks, and mean SSIM on the enhancement task are reported to compare SimMIM-US, USFM-woS, USFM-woF, and USFM.

performance when our proposed frequency domain masking or organ-balanced sampling is eliminated from USFM. Due to the removal of organ-balanced sampling, USFM-woS biased the majority organ (Breast, 91.74%), overfitting meaningless Breast-specific structural details and ignoring the minority organs in the 3M-US database, such as the Thyroid (3.29%), Fetus (0.93%), Neck (0.51%) and Abdomen (0.43%). The lack of organ applicability in USFM-woS resulted in degraded performance than USFM on all organs, especially in the Breast (-3.2%)

and Thyroid (-1.4%) on segmentation, Fetus (-6.4%) and Abdomen (-6.4%) on classification, and the Carotid (-2.1%) and Liver (-2.4%) on image enhancement. Even biased towards Breast, USFM-woS does not perform better than USFM on downstream tasks of Breast, with -3.2% and -5.2% degradation on the tumor segmentation and classification. This result indicates that the overfitted Breast features in USFM-woS are not generalizable to the downstream task. USFM-woF, which removed the frequency band-stop masked learning, focused

solely on spatial domain information. Caused by the neglect of implicit frequency information, USFM-woF is insufficient to extract effective US features, thus leading to inferior performance compared to USFM. The inferior was more significant in the classification and enhancement tasks **Fig. 11** (b, c), which require a high understanding of the US image. The performance is degraded by -3.9%, -5%, and -3.1% in the classification of Breast tumors, Abdominal structures, and Fetal planes, respectively, and by -1.6%, -0.4%, -2.6%, and -1.6% in the enhancement of the Liver, kidneys, Thyroid, and Carotid, respectively.

As for the SimMIM-US, the comparison results show that SimMIM-US performs inferior to USFM on segmentation tasks (DSCs of 79.6%, 83.2%, 80.9%, 84.3% vs. 80.6%, 85.9%, 84.3%, 85.8% on Neck, Thyroid, Breast, and Fetus). SimMIM-US also performed significantly worse than USFM on classification and enhancement tasks across all organs. The superior performance of USFM demonstrates the effectiveness of its specific design for US images compared to SimMIM-US.

In contrast to SimMIM-US, USFM-woS and USFM-woF, USFM, by extracting organ-unbiased and effective US features, achieved the best performance across all organs in all tasks. The experimental results further validate the effectiveness of our proposed organ-balanced sampling and spatial-frequency dual masking method in MIM.

6. Discussion

The safety imaging principles and the versatile devices make US imaging more widely accessible for disease screening and health management than other medical imaging modalities. In hospital and community healthcare settings, numerous US images are produced in various organs and diseases and require automated analysis. Existing US analysis methods struggle to meet these needs in complex medical scenarios due to insufficient organ applicability, task adaptation, and label efficiency.

USFM, the US foundation model we have built in this study, offers a promising solution to address the shortcomings of current US models and meet the demands in clinical practice. In massive experiments, USFM has demonstrated strong cross-organ applicability, various task adaptability, and label efficiency. As shown in **Tables 2, 3, and 4**, USFM exhibits robust performance, significantly surpassing representative methods across various organs (Neck, Thyroid, Breast, Carotid, Liver, Kidney, Abdomen, and Fetus) and common US tasks (segmentation, classification, and image enhancement). Moreover, USFM shows exceptional label efficiency in segmentation and classification tasks, as illustrated in **Figs. 9** and **10**. Notably, even with only 20% annotated data, the performance of USFM leads ahead of existing methods trained on complete datasets (superior to UperNet on Neck and Thyroid segmentation, superior to Segformer on Thyroid and Breast segmentation). These excellent properties of USFM stem from our innovative and practical approach to the foundation model construction, addressing three long-standing challenges in the US image: insufficient databases, low quality, and feature ineffectiveness.

Current US models are generally developed on an insufficient database, consisting of images of a particular organ from a single center using the same device. The lack of diversity in training data makes these models unsuitable for complex real-world scenarios, where the US image datasets exhibit significant variability among organs, operator inconsistencies, and equipment deviations. To address this problem, we have established the largest Multi-organ, Multi-center, Multi-device US database, 3M-US, containing over two million US images to cover as many organs, diseases, and centers as possible. Through full training on the large-scale database, USFM is universally applicable to many clinical scenarios. As demonstrated in **Table 2**, compared to UperNet with the same network structure but without 3M-US pre-training, USFM shows superior universality across all downstream tasks for all organs. Although SimMIM, pre-trained on natural images, can capture some structural information shared with US images, it fails to extract

high-level semantic US features. Therefore, SimMIM consistently underperforms USFM, particularly in classification and enhancement tasks that require a deep understanding of US images. These comparisons validate the comprehensive variety of our 3M-US database, serving as a data basis for establishing a universal US foundation model.

The superior performance of USFM further demonstrates the relationship between the universality and effectiveness of foundation models. To learn organ-universal features, we employed an organ-balanced sampling strategy in USFM. By balanced learning from different organs in the 3M-US dataset, USFM has adequate feature extraction and reconstruction capabilities for all 12 organs, as demonstrated in **Fig. 6**. The effectiveness of these organ-universal features is evidenced in ablation experiments, as shown in **Fig. 11**. USFM-woS, removing organ-balanced sampling from USFM, leads to overemphasizing the dominant organs (Breast) during pre-training. The Breast-biased USFM-woS, collapsed to extract Breast features in US images like edges and calcifications, fails in the other organs. This failure significantly degrades performance across all tasks and organs compared to USFM, particularly in classification and enhancement tasks requiring more useful high-level US features. Even on breast tumor segmentation and classification, the organ-universal USFM is more effective than the Breast-biased USFM-woS. The pursuit of universality ultimately leads to enhanced effectiveness, highlighting the significance of developing universal foundation models.

The innovation of USFM also lies in the specially designed spatial-frequency dual masked MIM for feature learning on low-quality US images. The main challenge of low-quality US images is that the massive inherent noise submerges the helpful image signals, resulting in the failure of the model to extract meaningful US features. Existing US image denoising methods are limited by computation resources for large-scale applications. Instead of complex image denoising, this study proposes a novel spatial-frequency dual masked (noise-added) method to learn the features from image noise. In the dual masked MIM pre-training, The USFM is optimized to extract meaningful image signals to reconstruct the original US image from massive added artificial noises. As shown in **Fig. 6**, the USFM is robust to noise, proven by the similar SNR between the reconstructed US image and the original image, even severely masked by the artificial noise. Artificial noise is not added to unmasked US images in downstream tasks. Thus, the strong artificial noise-adapted USFM can robustly extract useful features from the insignificant inherent noise (compared to artificial noise) to achieve sufficient performance in the downstream tasks. As demonstrated in **Table 2** and **Fig. 8**, USFM obtains superior and stable performance compared to other methods (higher mean and lower std on all metrics). In addition, robustness to noise also allows USFM to learn efficiently from limited data, as the best label efficiency is shown in **Fig. 10**. In addition, robustness to noise also allows USFM to learn efficiently from limited data, as the best label efficiency is shown in **Fig. 10**. Our proposed spatial-frequency dual masked MIM approach offers a new direction for future research on low-quality US images.

The inefficiencies of US features are attributed to the scarcity of spatial information and the difficulty of useful knowledge extraction. In the spatial domain, only limited local information is available for US model extraction due to low resolution, low SNR, and low contrast. Concurrently, tasks such as the classification of benign and malignant Breast tumors require a high-level understanding of rapidly changing texture details in US images, which are challenging to capture in the spatial domain. Our USFM employs a frequency band-stop masking learning method within MIM to represent this useful information directly. The frequency domain reconstruction visualization in **Fig. 7** shows that USFM can extract useful frequency information from the residual components. The ability to capture frequency domain information allows it to significantly outperform the spatial-only learning USFM-woF across all tasks and organs, as proven in the ablation experiment in **Fig. 11**. Especially in classification and enhancement tasks, the

Table 5

Comparison with the task-specific model on the downstream tasks.

Organs	Models	Task-specific design	Segmentation		
			DSC (%)	IoU (%)	
Neck (Nerve)	Van Boxtel et al. (2021) USFM	Ensemble	83.0	70.9	
		–	80.6	69.5	
Thyroid (TN3K)	Gong et al. (2023) USFM	Dataset	83.3	71.4	
		–	85.9	77.5	
Breast (BUSI)	Lin et al. (2023) USFM	Prompts	85.8	75.1	
		–	84.3	76.0	
			Classification		
			ACC (%)	Recall (%)	
Breast (BUSI)	Ghefati and Rivaz (2022) USFM	Ensemble	88.9	–	
		–	87.7	84.4	
Abdomen (USAnotAI)	Li et al. (2021) USFM	Algorithm	96.3	–	
		–	93.3	93.3	
Fetus (Fetal Planes)	Burgos-Artiz et al. (2020) USFM	Network	93.7	92.5	
		–	93.6	94.0	

neglect of frequency information in USFM-woF cannot extract higher-level semantic features, leading to a substantial decline in performance. By combining spatial mean masking learning and frequency band-stop masking learning, our dual-domain masked MIM approach enables USFM to achieve superior performance and high label efficiency in various tasks and organs, as demonstrated in Tables 3, 4, and Fig. 10.

On the downstream tasks involved in USFM, many studies have established the task-specific model with SOTA performance by unique design, such as incorporating additional data (Gong et al., 2023), model ensemble (Van Boxtel et al., 2021; Ghefati and Rivaz, 2022), and promoting with the label (Lin et al., 2023). The comparison of the foundation model (USFM) and task-specific models will provide a view of the relationship between the two modeling routes in medical image analysis. As presented in Table 5, the foundation model USFM achieved similar performance with the task-specific model on the segmentation and classification tasks. On the Neck and Breast segmentation tasks, the task-specific model achieved slightly higher performance than the USFM by ensemble multiple models and label prompting (DSCs of 83.0%, 85.8% vs. 80.6%, 84.3%). On the Thyroid segmentation task, USFM performs better by pre-training on a larger US dataset (DSCs of 85.9% vs. 83.9%). In both classification tasks, the difference in F1 scores between the USFM and the SOTA task-specific models is less than 3% (F1 scores of 87.7%, 93.3%, 93.6% vs. 88.9%, 96.3%, 93.7%). The similar performance of the USFM and task-specific model indicates that the two routes are intersecting with the development of AI. The task-universal data knowledge learned by the USFM on the large-scale dataset plays an equally important role in downstream tasks as the specific design from the prior task knowledge. The foundation model is promising for rapid adaptation on multiple downstream medical tasks with comparable performance to task-specific models without additional complex task-specific design.

While our USFM has achieved excellent performance, there is room for improvement. Firstly, our study primarily focuses on 2D US images. Although the analysis of 3D US images can be converted to 2D images, directly establishing a foundation model for 3D US would better utilize the relationships between images. Secondly, limited by computational resources, the potential performance improvements of larger foundation model architectures (ViT-L, ViT-H, Swin Transformer Liu et al., 2021) remain to be explored. In addition to segmentation, classification, and image enhancement tasks conducted in experiments, USFM also has the potential to be applied to a broader range of US image analysis tasks, such as detection, denoising, and generation. In the future, we will continue exploring effective foundation modeling methods for US images. Moreover, the rapid adaptation of US foundation models to few-shot downstream tasks with prompt learning is a future direction for research to expand the foundation model application scope and achieve widespread use in clinical practice.

7. Conclusion

This paper developed a medical US image foundation model named USFM, characterized by high organ versatility, task adaptability, and label efficiency via sufficient pre-training on the 3M-US database. Comprehensive experiments demonstrate that USFM exhibits solid performance in various common US tasks across different organs, including segmentation, classification, and enhancement. The outstanding performance of USFM in label efficiency experiments confirms its potential as a foundation model to accelerate the development of US models and promote the expansion of automated US analysis applications.

CRediT authorship contribution statement

Jing Jiao: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Project administration, Methodology, Data curation, Conceptualization. **Jin Zhou:** Data curation. **Xiaokang Li:** Writing – review & editing. **Menghua Xia:** Writing – review & editing. **Yi Huang:** Investigation. **Lihong Huang:** Writing – original draft. **Na Wang:** Investigation. **Xiaofan Zhang:** Project administration. **Shichong Zhou:** Data curation. **Yuanyuan Wang:** Writing – review & editing, Supervision, Project administration. **Yi Guo:** Writing – review & editing, Project administration.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant 62371139 and 82227803), the Science and Technology Commission of Shanghai Municipality, China (Grant 22DZ1100100 and 22ZR1404800).

References

- Al-Dhabayani, W., Gomaa, M., Khaled, H., Fahmy, A., 2020. Dataset of breast ultrasound images. *Data Brief* 28, 104863. <http://dx.doi.org/10.1016/j.dib.2019.104863>.
- Antropova, N., Huynh, B.Q., Giger, M.L., 2017. A deep feature fusion methodology for breast cancer diagnosis demonstrated on three imaging modality datasets. *Med. Phys.* 44 (10), 5162–5171. <http://dx.doi.org/10.1002/mp.12453>.
- Awais, M., Naseer, M., Khan, S., Anwer, R.M., Cholakkal, H., Shah, M., Yang, M.-H., Khan, F.S., 2023. Foundational models defining a new era in vision: A survey and outlook. <arXiv:2307.13721>.
- Azad, B., Azad, R., Eskandari, S., Bozorgpour, A., Kazerouni, A., Rekik, I., Merhof, D., 2023. Foundational models in medical imaging: A comprehensive survey and future vision. <arXiv:2310.18689>.
- Bao, H., Dong, L., Piao, S., Wei, F., 2022. BEiT: BERT pre-training of image transformers. <http://dx.doi.org/10.48550/arXiv.2106.08254>, <arXiv:2106.08254>.
- Burgos-Artiz, X.P., Coronado-Gutiérrez, D., Valenzuela-Alcaraz, B., Bonet-Carne, E., Eixarch, E., Crispí, F., Gratacós, E., 2020. Evaluation of deep convolutional neural networks for automatic classification of common maternal fetal ultrasound planes. *Sci. Rep.* 10 (1), 10200. <http://dx.doi.org/10.1038/s41598-020-67076-5>.
- Chen, X., Fan, H., Girshick, R., He, K., 2020b. Improved baselines with momentum contrastive learning. <http://dx.doi.org/10.48550/arXiv.2003.04297>, <arXiv:2003.04297>.
- Chen, T., Kornblith, S., Norouzi, M., Hinton, G., 2020a. A simple framework for contrastive learning of visual representations. <http://dx.doi.org/10.48550/arXiv.2002.05709>, <arXiv:2002.05709>.
- Chen, T., Saxena, S., Li, L., Lin, T.-Y., Fleet, D.J., Hinton, G., 2022. A unified sequence interface for vision tasks. <arXiv:2206.07669>.
- Da Correggio, K.S., Noya Galluzzo, R., Santos, L.O., Soares Muylaert Barroso, F., Zimmermann Loureiro Chaves, T., Sherley Casimiro Onofre, A., von Wangenheim, A., Fetal abdominal structures segmentation dataset using ultrasonic images 1. <http://dx.doi.org/10.17632/4gcpm9dsc3.1>.

- Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N., 2021. An image is worth 16x16 words: Transformers for image recognition at scale. <http://dx.doi.org/10.48550/arXiv.2010.11929>, arXiv:2010.11929.
- Fontanarosa, D., van der Meer, S., Bamber, J., Harris, E., O'Shea, T., Verhaegen, F., 2015. Review of ultrasound image guidance in external beam radiotherapy: I. Treatment planning and inter-fraction motion management. *Phys. Med. Biol.* 60 (3), R77. <http://dx.doi.org/10.1088/0031-9155/60/3/R77>.
- Ghefati, B., Rivaz, H., 2022. Vision transformers for classification of breast ultrasound images. In: 2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society. EMBC, pp. 480–483. <http://dx.doi.org/10.1109/EMBC48229.2022.9871809>.
- Ghojogh, B., Crowley, M., Karray, F., Ghodsi, A., 2023. Uniform manifold approximation and projection (UMAP). In: Ghojogh, B., Crowley, M., Karray, F., Ghodsi, A. (Eds.), *Elements of Dimensionality Reduction and Manifold Learning*. Springer International Publishing, Cham, pp. 479–497. http://dx.doi.org/10.1007/978-3-031-10602-6_17.
- Gong, H., Chen, J., Chen, G., Li, H., Li, G., Chen, F., 2023. Thyroid region prior guided attention for ultrasound segmentation of thyroid nodules. *Comput. Biol. Med.* 155, 106389. <http://dx.doi.org/10.1016/j.combiomed.2022.106389>.
- Guo, Y., Huang, L., Xia, M., Zhou, Z., Jiao, J., 2023. Ultrasound image enhancement challenge 2023. URL: <https://ultrasoundenhance2023.grand-challenge.org.ultrasoundenhance2023/>.
- He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R., 2022. Masked autoencoders are scalable vision learners. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR, pp. 15979–15988. <http://dx.doi.org/10.1109/CVPR52688.2022.01553>.
- He, K., Fan, H., Wu, Y., Xie, S., Girshick, R., 2020. Momentum contrast for unsupervised visual representation learning. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR, pp. 9726–9735. <http://dx.doi.org/10.1109/CVPR42600.2020.00975>.
- He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition. CVPR, pp. 770–778. <http://dx.doi.org/10.1109/CVPR.2016.90>.
- Isensee, F., Jaeger, P.F., Kohl, S.A.A., Petersen, J., Maier-Hein, K.H., 2021. nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods* 18 (2), 203–211. <http://dx.doi.org/10.1038/s41592-020-01008-z>.
- Jiang, L., Dai, B., Wu, W., Loy, C.C., 2021. Focal frequency loss for image reconstruction and synthesis. In: 2021 IEEE/CVF International Conference on Computer Vision. ICCV, pp. 13899–13909. <http://dx.doi.org/10.1109/ICCV48922.2021.01366>.
- Kang, Q., Gao, J., Li, K., Lao, Q., 2023. Deblurring masked autoencoder is better recipe for ultrasound image recognition. In: Greenspan, H., Madabhushi, A., Mousavi, P., Salcudean, S., Duncan, J., Syeda-Mahmood, T., Taylor, R. (Eds.), *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*. Springer Nature Switzerland, Cham, pp. 352–362. http://dx.doi.org/10.1007/978-3-031-43907-0_34.
- Kang, Q., Lao, Q., Li, Y., Jiang, Z., Qiu, Y., Zhang, S., Li, K., 2022. Thyroid nodule segmentation and classification in ultrasound images through intra- and inter-task consistent learning. *Med. Image Anal.* 79, 102443. <http://dx.doi.org/10.1016/j.media.2022.102443>.
- Kim-Ann, 2020. “Organ classification on abdominal ultrasound using javascript. <https://github.com/ftsvd/USAnotAI>.
- Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.-Y., Dollár, P., Girshick, R., 2023. Segment anything. <http://dx.doi.org/10.48550/arXiv.2304.02643>, arXiv:2304.02643.
- Lei, W., Su, Q., Jiang, T., Gu, R., Wang, N., Liu, X., Wang, G., Zhang, X., Zhang, S., 2024. One-shot weakly-supervised segmentation in 3D medical images. *IEEE Trans. Med. Imaging* 43 (1), 175–189. <http://dx.doi.org/10.1109/TMI.2023.3294975>.
- Li, T., Chang, H., Mishra, S.K., Zhang, H., Katabi, D., Krishnan, D., 2023. MAGE: MAsked generative encoder to unify representation learning and image synthesis. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR, pp. 2142–2152. <http://dx.doi.org/10.1109/CVPR52729.2023.00213>.
- Li, K., Xu, Y., Meng, M.Q.-H., 2021. Automatic recognition of abdominal organs in ultrasound images based on deep neural networks and k-nearest-neighbor classification. arXiv:2110.04563.
- Lin, X., Xiang, Y., Zhang, L., Yang, X., Yan, Z., Yu, L., 2023. SAMUS: Adapting segment anything model for clinically-friendly and generalizable ultrasound image segmentation. arXiv:2309.06824.
- Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B., 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In: 2021 IEEE/CVF International Conference on Computer Vision. ICCV, pp. 9992–10002. <http://dx.doi.org/10.1109/ICCV48922.2021.00986>.
- Liu, S., Wang, Y., Yang, X., Lei, B., Liu, L., Li, S.X., Ni, D., Wang, T., 2019. Deep learning in medical ultrasound analysis: A review. *Engineering* 5 (2), 261–275. <http://dx.doi.org/10.1016/j.eng.2018.11.020>.
- Lu, C.-Z., Jin, X., Hou, Q., Liew, J.H., Cheng, M.-M., Feng, J., 2023. Delving deeper into data scaling in masked image modeling. arXiv:2305.15248.
- Luo, X., Liao, W., Xiao, J., Chen, J., Song, T., Zhang, X., Li, K., Metaxas, D.N., Wang, G., Zhang, S., 2022. WORD: A large scale dataset, benchmark and clinical applicable study for abdominal organ segmentation from CT image. *Med. Image Anal.* 82, 102642. <http://dx.doi.org/10.1016/j.media.2022.102642>.
- Montoya, A., Hasnain, k.s.W.C.y., 2016. Ultrasound nerve segmentation. URL: <https://kaggle.com/competitions/ultrasound-nerve-segmentation>.
- Qie, Y., Lee, H.J., 2018. Densely connected convolutional networks for multi-exposure fusion. In: 2018 International Conference on Computational Science and Computational Intelligence. CSCI, pp. 417–420. <http://dx.doi.org/10.1109/CSCI46756.2018.00084>.
- Safa, A., Mohamed, A., Issam, B., Mohamed-Yassine, H., 2023. SegFormer: Semantic segmentation based tranformers for corrosion detection. In: 2023 International Conference on Networking and Advanced Systems. ICNAS, pp. 1–6. <http://dx.doi.org/10.1109/ICNAS59892.2023.10330461>.
- Van Boxtel, J., Vousten, V., Pluim, J., Rad, N.M., 2021. Hybrid deep neural network for brachial plexus nerve segmentation in ultrasound images. In: 2021 29th European Signal Processing Conference. EUSIPCO, pp. 1246–1250. <http://dx.doi.org/10.23919/EUSIPCO54536.2021.9616329>.
- Wang, Z., Liu, C., Zhang, S., Dou, Q., 2023c. Foundation model for endoscopy video analysis via large-scale self-supervised pre-train. In: Greenspan, H., Madabhushi, A., Mousavi, P., Salcudean, S., Duncan, J., Syeda-Mahmood, T., Taylor, R. (Eds.), *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*. Springer Nature Switzerland, Cham, pp. 101–111. http://dx.doi.org/10.1007/978-3-031-43996-4_10.
- Wang, D., Wang, X., Wang, L., Li, M., Da, Q., Liu, X., Gao, X., Shen, J., He, J., Shen, T., Duan, Q., Zhao, J., Li, K., Qiao, Y., Zhang, S., 2023a. A real-world dataset and benchmark for foundation model adaptation in medical image classification. *Sci. Data* 10 (1), 574. <http://dx.doi.org/10.1038/s41597-023-02460-0>.
- Wang, G., Wu, J., Luo, X., Liu, X., Li, K., Zhang, S., 2023b. MIS-FM: 3D medical image segmentation using foundation models pretrained on a large-scale unannotated dataset. <http://dx.doi.org/10.48550/arXiv.2306.16925>, arXiv:2306.16925.
- Wei, Y., Hu, H., Xie, Z., Zhang, Z., Cao, Y., Bao, J., Chen, D., Guo, B., 2022. Contrastive learning rivals masked image modeling in fine-tuning via feature distillation. <http://dx.doi.org/10.48550/arXiv.2205.14141>, arXiv:2205.14141.
- Wu, L., Gao, X., Hu, Z., Zhang, S., 2024. Pattern-aware transformer: Hierarchical pattern propagation in sequential medical images. *IEEE Trans. Med. Imaging* 43 (1), 405–415. <http://dx.doi.org/10.1109/TMI.2023.3306468>.
- Xia, M., Yang, H., Qu, Y., Guo, Y., Zhou, G., Zhang, F., Wang, Y., 2022. Multi-level structure-preserved GAN for domain adaptation in intravascular ultrasound analysis. *Med. Image Anal.* 82, 102614. <http://dx.doi.org/10.1016/j.media.2022.102614>.
- Xiao, T., Liu, Y., Zhou, B., Jiang, Y., Sun, J., 2018. Unified perceptual parsing for scene understanding. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (Eds.), *Computer Vision – ECCV 2018*. Springer International Publishing, Cham, pp. 432–448. http://dx.doi.org/10.1007/978-3-030-01228-1_26.
- Xie, Z., Zhang, Z., Cao, Y., Lin, Y., Bao, J., Yao, Z., Dai, Q., Hu, H., 2022. SimMIM: A simple framework for masked image modeling. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR, pp. 9643–9653. <http://dx.doi.org/10.1109/CVPR52688.2022.00943>.
- Yuan, L., Chen, D., Chen, Y.-L., Codella, N., Dai, X., Gao, J., Hu, H., Huang, X., Li, B., Li, C., Liu, C., Liu, M., Liu, Z., Lu, Y., Shi, Y., Wang, L., Wang, J., Xiao, B., Xiao, Z., Yang, J., Zeng, M., Zhou, L., Zhang, P., 2021. Florence: A new foundation model for computer vision. <http://dx.doi.org/10.48550/arXiv.2111.11432>, arXiv:2111.11432.
- Zhang, Y., Gao, J., Tan, Z., Zhou, L., Ding, K., Zhou, M., Zhang, S., Wang, D., 2024. Data-centric foundation models in computational healthcare: A survey. <http://dx.doi.org/10.48550/arXiv.2401.02458>, arXiv:2401.02458.
- Zhang, Z., Liu, Q., Wang, Y., 2018. Road extraction by deep residual U-net. *IEEE Geosci. Remote Sens. Lett.* 15 (5), 749–753. <http://dx.doi.org/10.1109/LGRS.2018.2802944>.
- Zhang, S., Metaxas, D., 2024. On the challenges and perspectives of foundation models for medical image analysis. *Med. Image Anal.* 91, 102996. <http://dx.doi.org/10.1016/j.media.2023.102996>.
- Zhou, Y., Chia, M.A., Wagner, S.K., Ayhan, M.S., Williamson, D.J., Struyven, R.R., Liu, T., Xu, M., Lozano, M.G., Woodward-Court, P., Kihara, Y., Altmann, A., Lee, A.Y., Topol, E.J., Denniston, A.K., Alexander, D.C., Keane, P.A., 2023. A foundation model for generalizable disease detection from retinal images. *Nature* 622 (7981), 156–163. <http://dx.doi.org/10.1038/s41586-023-06555-x>.
- Zhu, J.-Y., Park, T., Isola, P., Efros, A.A., 2017. Unpaired image-to-image translation using cycle-consistent adversarial networkss. In: *Computer Vision (ICCV), 2017 IEEE International Conference on*.