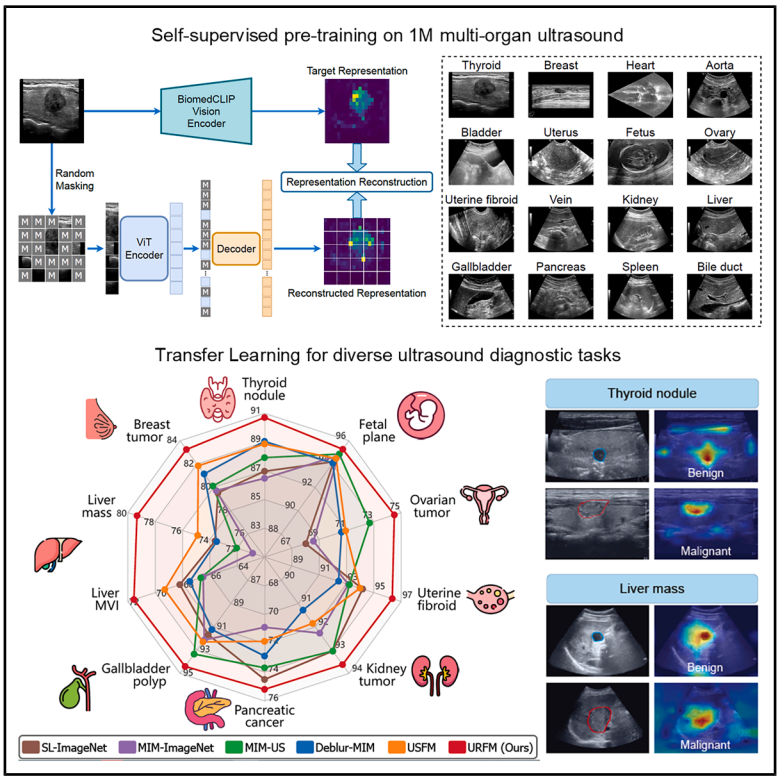
iScience

Article

URFM: A general Ultrasound Representation Foundation Model for advancing ultrasound image diagnosis

Graphical abstract



Highlights

- A universal ultrasound foundation model pre-trained on over 1 million images
- Uses representation-based masked image modeling to handle low SNR ultrasound imaging
- Leverages high-level medical vision-language features to enhance representation
- Achieves state-of-the-art performance across diverse ultrasound diagnostic tasks

Authors

Qingbo Kang, Qicheng Lao, Jun Gao, ..., Chenlin Du, Qiang Lu, Kang Li

Correspondence

qicheng.lao@bupt.edu.cn (Q.L.), luqiang@scu.edu.cn (Q.L.), likang@wchscu.cn (K.L.)

In brief

Health sciences; Computer-aided diagnosis method; Ultrasound technology; Computer science; Artificial intelligence; Machine learning

Kang et al., 2025, iScience 28, 112917
August 15, 2025
https://doi.org/10.1016/j.isci.2025.112917

CellPress

iScience

CellPress
OPEN ACCESS

Article

URFM: A general Ultrasound Representation Foundation Model for advancing ultrasound image diagnosis

Qingbo Kang,[1,4,5] Qicheng Lao,[2,*] Jun Gao,[1,4,5,7] Wuyongga Bao,[3] Zhu He,[2] Chenlin Du,[8] Qiang Lu,[3,*] and Kang Li[1,4,5,6,9,*]

[1]West China Biomedical Big Data Center, West China Hospital, Sichuan University, Chengdu 610041, Sichuan, China
[2]School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing 100876, China
[3]Department of Ultrasonography, West China Hospital, Sichuan University, Chengdu 610041, Sichuan, China
[4]Med-X Center for Informatics, Sichuan University, Chengdu, Sichuan 610041, China
[5]West China Hospital-SenseTime Joint Lab, Chengdu, Sichuan 610041, China
[6]Sichuan University - Pittsburgh Institute, Sichuan University, Chengdu, Sichuan 610207, China
[7]Stork Healthcare, Chengdu 610041, Sichuan, China
[8]School of Biomedical Engineering, Tsinghua University, Beijing 100084, China
[9]Lead contact
*Correspondence: qicheng.lao@bupt.edu.cn (Q.L.), luqiang@scu.edu.cn (Q.L.), likang@wchscu.cn (K.L.)
https://doi.org/10.1016/j.isci.2025.112917

SUMMARY

Ultrasound imaging is critical for clinical diagnostics, providing insights into various diseases and organs. However, artificial intelligence (AI) in this field faces challenges, such as the need for large labeled datasets and limited task-specific model applicability, particularly due to ultrasound's low signal-to-noise ratio (SNR). To overcome these, we introduce the Ultrasound Representation Foundation Model (URFM), designed to learn robust, generalizable representations from unlabeled ultrasound images, enabling label-efficient adaptation to diverse diagnostic tasks. URFM is pre-trained on over 1M images from 15 major anatomical organs using representation-based masked image modeling (MIM), an advanced self-supervised learning. Unlike traditional pixel-based MIM, URFM integrates high-level representations from BiomedCLIP, a specialized medical vision-language model, to address the low SNR issue. Extensive evaluation shows that URFM outperforms state-of-the-art methods, offering enhanced generalization, label efficiency, and training-time efficiency. URFM's scalability and flexibility signal a significant advancement in diagnostic accuracy and clinical workflow optimization in ultrasound imaging.

INTRODUCTION

Ultrasound, a foundational imaging modality in medicine, is extensively utilized for diagnosing a wide range of diseases across multiple anatomical organs due to its real-time capabilities, low cost, convenience, non-invasiveness, and lack of radiation exposure.[1,2] AI techniques, particularly deep learning (DL), have been widely adopted in ultrasound imaging, significantly enhancing diagnostic accuracy.[3–10] However, these AI models typically rely on extensive labeled datasets trained through supervised learning, making the acquisition of high-quality labeled data costly and labor-intensive. Furthermore, such task-specific models often have a narrow focus, limiting their generalizability across diverse clinical scenarios.[11,12] Last, but not least, a key challenge unique to ultrasound imaging, compared to natural images and other medical imaging modalities, is its inherently low signal-to-noise ratio (SNR), which may complicate AI model performance.

To mitigate the reliance on labeled data and address the challenge of limited generalizability, the medical imaging field has increasingly embraced the pre-training and fine-tuning paradigm, leveraging large-scale pre-trained foundation models to achieve remarkable advancements.[13–16] Unlike traditional DL models, foundation models consist of two stages: pre-training on a large volume of unlabeled data and subsequent downstream fine-tuning on smaller labeled datasets for specific tasks using supervised learning. Numerous studies have illustrated that foundation models significantly enhance downstream task performance while substantially reducing the dependency on task-specific labeled data.[17–23] Given the diverse imaging modalities in medicine, foundation models are typically tailored to specific modalities. For instance, RETFound[17] targets retinal images, Prov-GigaPath[23] focuses on pathological images, and EVA-X[24] is built for chest X-ray. These models are generally pre-trained on large-scale unlabeled data through self-supervised learning (SSL).[25,26] The SSL-based pre-training, encompassing two key approaches —contrastive learning[27–33] and masked image modeling (MIM)[34–36]— enables foundation models to capture generalized, task-agnostic representations

from unlabeled data, which can then be fine-tuned for various downstream tasks using labeled data, thereby adapting their learned representations to specific applications.

Several pioneering studies have applied SSL to ultrasound imaging to develop ultrasound-specific foundation models, achieving promising results.[21,37,38] Kang et al.[37,38] integrated a deblurring task into the original mask-and-reconstruct proxy task of MIM,[36,39] demonstrating superior performance in thyroid ultrasound. Jiao et al.[21] proposed USFM, a multiorgan ultrasound foundation model pre-trained through spatial-frequency dual MIM. However, these studies have been limited to a single organ[37,38] or face challenges in generalization due to the extremely severe imbalance in the pre-training dataset, with over 91% of the data originating from breast ultrasound, potentially limiting its adaptability to other organs and clinical scenarios.[21] Of particular importance, ultrasound inherently exhibits a low SNR due to its acoustic imaging process, complicating the effectiveness of current SSL approaches that predominantly rely on low-level pixel representations, which are unable to capture meaningful features from noisy and heterogeneous ultrasound data. Furthermore, a comprehensive evaluation across multiple organ systems and pathologies in the context of ultrasound-based clinical applications remains notably absent from current literature.[21,38] In conclusion, existing ultrasound foundation models are constrained by limited data diversity and size for pre-training, as well as insufficient evaluation of generalization across varied ultrasound applications. Addressing these limitations is essential for advancing foundation models that generalize effectively and transfer seamlessly to real-world clinical settings.

To address these limitations, we introduce URFM, an **U**ltrasound **R**epresentation **F**oundation **M**odel developed through representation-based MIM pre-training on a million-scale multi-organ ultrasound dataset. Figure 1 provides an overview of our study. We first construct a large-scale unlabeled ultrasound dataset for pre-training, comprising 1,003,448 images across 15 major anatomical organs or body parts commonly examined via ultrasound. The detailed organ distribution and data distribution by organ are illustrated in Figure 1A, demonstrating a relatively balanced distribution across different organs.

Following pre-training dataset construction, we adopt a representation-based MIM approach for URFM pre-training (Figure 1B). Given the inherently low SNR in ultrasound imaging, traditional pixel-based SSL methods are unsuitable, as they struggle to extract meaningful feature representations from the noisy data. Instead of pixel reconstruction in the original MIM framework, our method focuses on representation reconstruction using high-level semantic representations provided by the domain-specific BiomedCLIP.[40] This approach can also be viewed as a form of knowledge distillation, allowing URFM to integrate intrinsic medical insights during pre-training and achieve superior downstream performance. Our ablation results further verify that BiomedCLIP representations outperform both raw pixel-level targets and representations derived from general vision-language models such as CLIP.[41]

Finally, we evaluate URFM through supervised downstream fine-tuning across 10,000 clinical applications spanning 9 organs and varying diagnostic difficulties, such as the diagnosis of thyroid nodules,[42] liver masses,[7] liver microvascular invasion (MVI), and pancreatic cancer (Figures 1C and 1D)[8]. Experimental results (Figure 2) demonstrate that URFM achieves state-of-the-art (SOTA) performance across all downstream tasks, surpassing previous foundation models.[21,36,38,43] For instance, URFM improves the F1 score for liver mass diagnosis from the current SOTA of 74.89% to a new SOTA of 79.33%, liver MVI diagnosis from 69.3% to 71.56%, and ovarian tumor diagnosis from 70.94% to 74.5%. In addition, URFM exhibits superior label efficiency and reduced training time during fine-tuning.

The superiority of URFM is attributed to two principal factors: the dataset and methodology. Although thyroid accounts for 44.25% of the dataset, the inclusion of 14 additional organs ensures sufficient anatomical diversity, preventing overfitting to thyroid-specific features. In addition, URFM employs representation-based MIM to reconstruct high-level semantics from BiomedCLIP representations rather than raw pixels, making it less sensitive to organ imbalance and more robust to low-SNR ultrasound characteristics. This is supported by the strong consistency between attention maps during pre-training and downstream fine-tuning, and by empirical results showing that balanced sampling brought negligible improvement. In summary, URFM demonstrates superior generalization capabilities across diverse anatomical organs in ultrasound, solidifying its position as a generic, versatile, and efficient ultrasound foundation model. This versatility makes URFM a powerful tool for advancing AI applications in ultrasound imaging.

RESULTS

We developed URFM, a general ultrasound foundation model pre-trained on a million-scale, diverse ultrasound image dataset using advanced SSL technique tailored specifically for ultrasound image analysis with an emphasis on ultrasound feature representations. As illustrated in (a), the pre-training dataset comprises 1,003,465 ultrasound images covering 15 anatomical organs, originating from multiple sources. More details regarding pre-training datasets can be found in Table 1. Given the intrinsic property of ultrasound imaging, specifically its low SNR in the pixel space, we designed a representation-based MIM approach. This involved using the representations extracted by medical foundation models such as BiomedCLIP[40] as the reconstruction target within the MIM framework to pre-train the URFM ((b)). The generalizability of URFM was evaluated through fine-tuning on 10 ultrasound applications in various clinical scenarios (Figures 1C and 1D), demonstrating its superiority compared to other pre-trained foundation models.

URFM achieves state-of-the-art performance across diverse ultrasound applications

To comprehensively evaluate the generalizability of URFM, we designed a benchmark evaluation consisting of ten distinct ultrasound clinical applications, addressing a spectrum of diagnostic targets across nine organs, varying in diagnostic difficulty. Specifically, as shown in Figures 1C and 1D, the ten applications include the diagnosis of thyroid nodule, breast tumor, liver mass, liver microvascular invasion, gallbladder polyp, pancreatic cancer, kidney tumor, uterine fibroid, ovarian tumor, and fetal

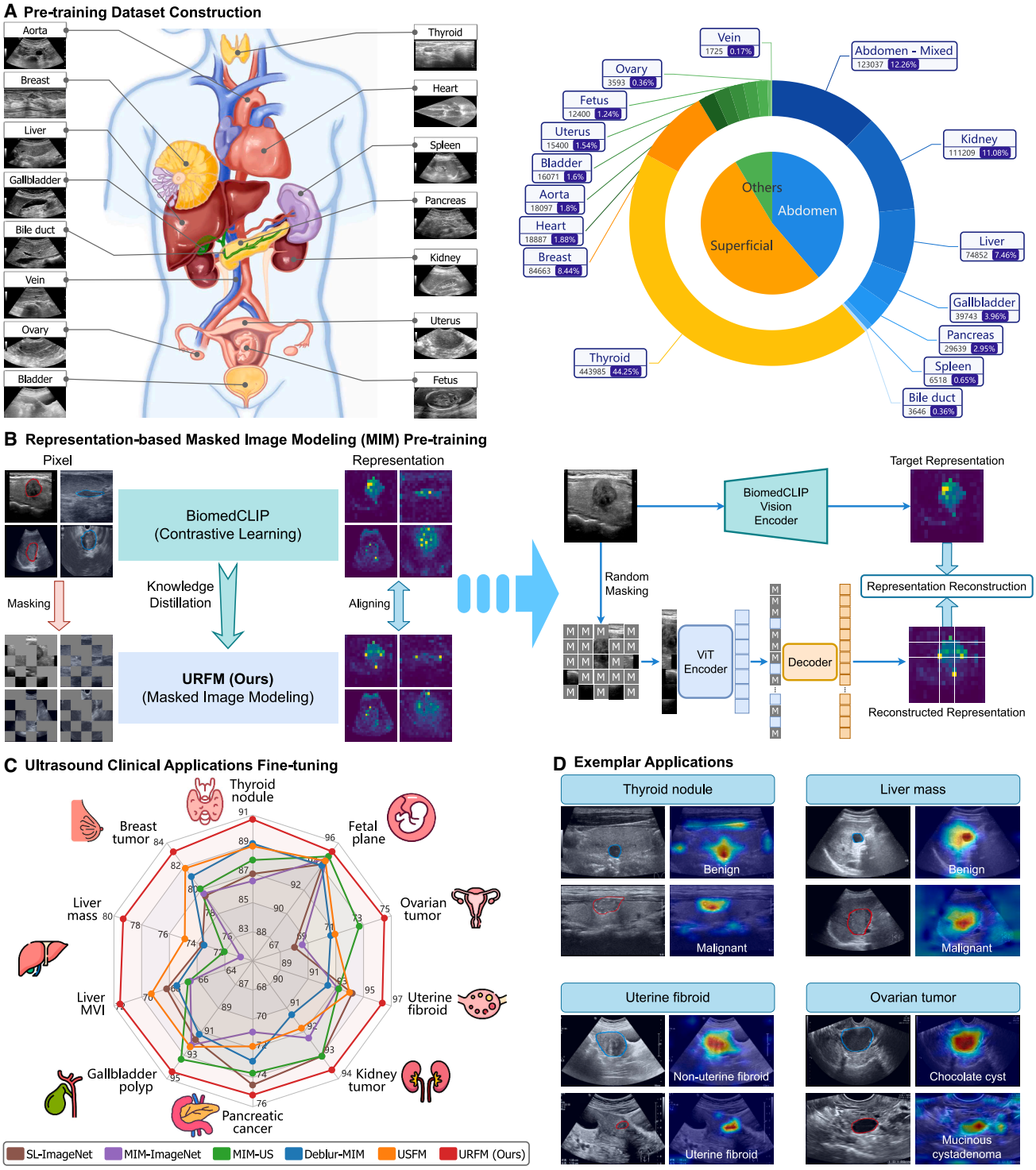


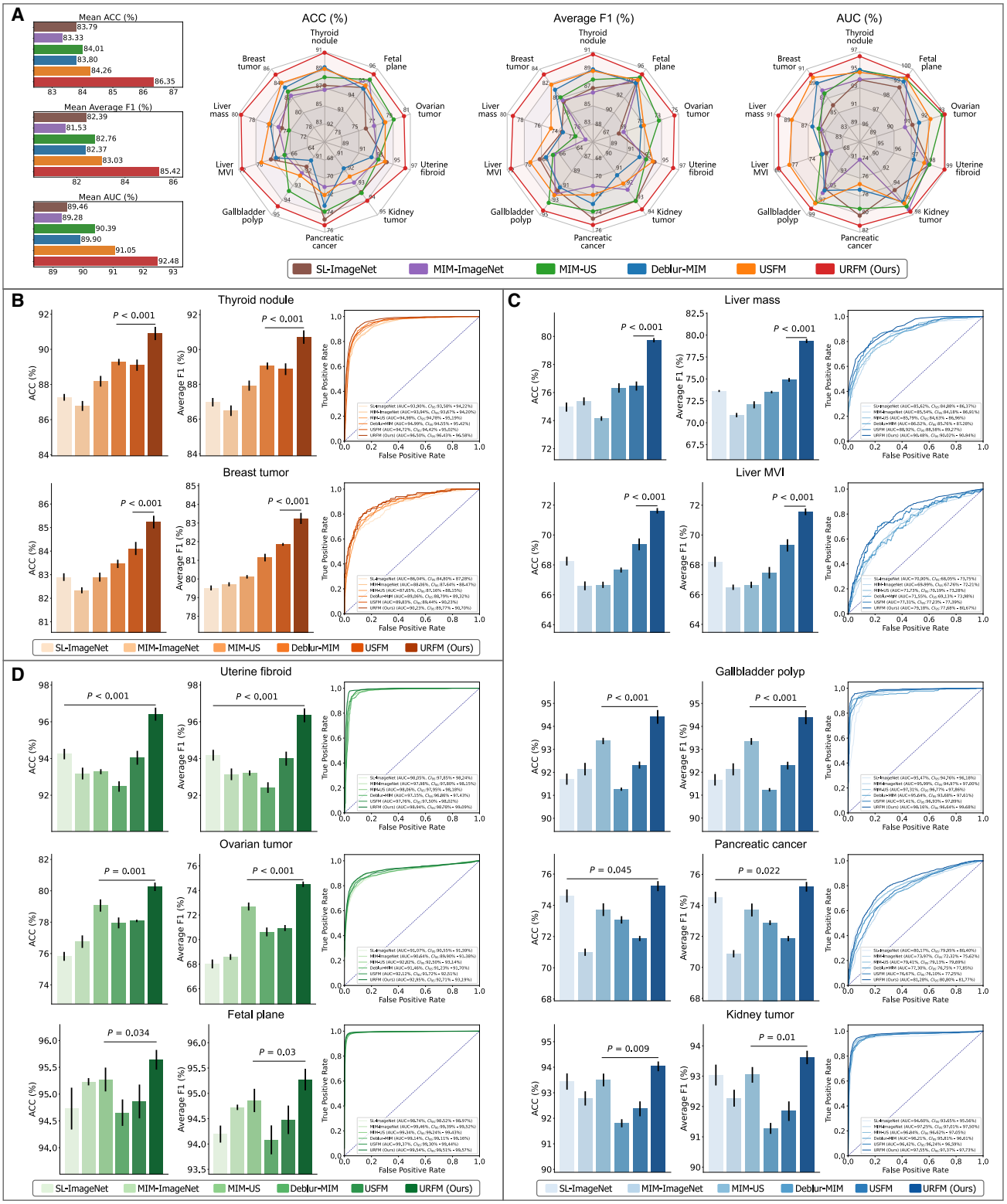
Figure 1. General overview of the study

(A) Pre-training dataset construction: we construct an ultrasound dataset comprising over 1 million images across 15 major anatomical organs.

(B) Representation-based MIM pre-training: we adopt reconstruct representations extracted by BiomedCLIP[40] within the MIM framework to pre-train URFM.

(C) Downstream fine-tuning: we perform comprehensive evaluation on ten ultrasound clinical applications across nine organs and varying diagnostic difficulties, the URFM achieves state-of-the-art performance across all downstream tasks, the performance is evaluated by average F1 metric.

(D) Visualizations for four exemplar downstream ultrasound applications.



(legend on next page)

plane. More details regarding these datasets can be found in Table 2 in the STAR Methods section.

We compare the performance of URFM against a range of existing foundation models, including those pre-trained via supervised learning, such as SL-ImageNet,[43] as well as SSL approaches, including MIM-ImageNet,[36] MIM-US, Deblur-MIM,[38] and the current state-of-the-art ultrasound foundation model, USFM.[21] The first two models are pre-trained on nature images (i.e., ImageNet[62]) with traditional supervised learning and the MIM approach,[36] respectively. The latter three models are pre-trained specifically on ultrasound images. MIM-US utilizes the dataset constructed in this study with the original MIM approach (i.e., pixel-based MIM[36]). Deblur-MIM is also pre-trained on our dataset, adopting the deblurring MIM proposed by Kang et al.[38] USFM is the model introduced by Jiao et al.[21] To ensure fairness, the network architecture for all models during both pre-training and downstream fine-tuning is the same ViT-Base/16 (ViT-B).[43]

Figure 2A presents comprehensive performance comparisons across all ten tasks, using three metrics: ACC (%), Average F1 (%), and the area under the ROC curve (AUC) (%). The three smaller bar charts in the figure show the mean performance metrics averaged across all tasks, while the three radar charts present detailed performance comparisons for each individual task. These averaged metrics clearly demonstrate that URFM outperforms all other models, achieving the best overall performance. For each task, URFM consistently ranks first across all three metrics. Specifically, URFM achieves a mean ACC of 86.35%, a mean Average F1 of 85.42%, and a mean AUC of 92.48%, outperforming the previous best results by 2.09%, 2.39%, and 1.43%, respectively. These comparisons highlight URFM's exceptional performance as a generic and versatile ultrasound foundation model, significantly enhancing clinical diagnosis across diverse anatomical organs. Furthermore, URFM's notable improvements over MIM-US highlight the effectiveness and superiority of representation-based MIM compared to traditional pixel-based MIM in ultrasound pre-training. Additionally, all models pre-trained on ultrasound images (URFM, USFM, Deblur-MIM, and MIM-US) outperform those pre-trained on nature images (SL-ImageNet and MIM-ImageNet), emphasizing the benefits of self-modality MIM pre-training on ultrasound imaging.

Next, we further detail the performance of URFM across various categories of ultrasound exams, including superficial organs (Figure 2B), abdominal organs (Figure 2C), and other organ systems (Figure 2D), respectively. We report three metrics: ACC (%) and Average F1 (%), together with receiver operating characteristic (ROC) curves for all ten tasks. The *P* values are calculated between URFM and the most competitive comparison model to check for significance.

Table 1. Details of the pre-training dataset

Organ	Source	Images	Subtotal
Thyroid	RadImageNet[44,a]	92,566	
	Private	351,419	443,985
Breast	BUV[45]	26,052	
	ultrasoundcases[b]	2,255	
	Private	56,356	84,663
Abdomen - Mixed	Private	123,037	123,037
Kidney	RadImageNet	111,209	111,209
Liver	RadImageNet	74,852	74,852
Gallbladder	RadImageNet	39,743	39,743
Pancreas	RadImageNet	21,639	
	LEPset[8]	8,000	29,639
Spleen	RadImageNet	6,518	6,518
Bile duct	RadImageNet	3,646	3,646
Heart	CAMUS[46]	18,887	18,887
Aorta	RadImageNet	18,097	18,097
Bladder	RadImageNet	758	
	Private	15,313	16,071
Uterus	RadImageNet	15,400	15,400
Fetus	FETAL_PLANES_DB[c]	12,400	12,400
Ovary	RadImageNet	3,593	3,593
Vein	RadImageNet	1,725	1,725

Our pre-training dataset comprises 1,003,465 images, covering 15 major anatomical organs or body parts, and is sourced from a variety of distinct origins.
[a]https://www.radimagenet.com.
[b]https://www.ultrasoundcases.info.
[c]https://zenodo.org/records/3904280.

Overall, URFM achieves the best performance in terms of all three metrics with statistical significance (all $P < 0.05$). For instance, for the breast tumor diagnosis task in Figure 2B, URFM achieves ACC = 85.23% (95% CI: 84.96%–85.50%), Average F1 = 83.25% (95% CI: 82.95%–83.54%), and AUC = 90.23% (95% CI: 89.77%–90.70%), which is significantly better (all < 0.001) than USFM, the second best-performing model. In another example, for the liver mass diagnosis task shown in Figure 2C, URFM outperforms all other models with ACC = 79.73% (95% CI: 79.57%–79.88%), Average F1 = 79.33% (95% CI: 79.09%–79.57%), and AUC = 90.48% (95% CI: 90.02%–90.94%), again demonstrating statistically significant improvements over the next best model (all $P < 0.001$).

To further assess URFM's reproducibility on external data, we conducted an evaluation on the publicly available BUID breast lesion ultrasound dataset.[49–51] As shown in Figure 3, URFM achieves an accuracy of 79.59%, an F1 score of 79.58%, and

Figure 2. Experimental results of the ten downstream ultrasound clinical applications are presented as follows
(A) Summary results of all ten tasks.
(B) Detailed results for tasks involving superficial organs.
(C) Detailed results for tasks involving abdominal organs.
(D) Detailed results for tasks involving other organs. Error bars represent the 95% confidence intervals (CI). Statistical significance is assessed using *P* values calculated from two-sided *t*-tests.

Table 2. The detailed dataset, classes, and number of images for each ultrasound clinical applications in downstream fine-tuning

Application	Dataset	Classes	Images
Thyroid nodule	Thyroid4K[4]	Benign, Malignant	4,493
Breast tumor[a]	BUS-BRA[47]	Benign, Malignant	1,064
Liver mass[b]	Liver735[7]	Benign, Malignant, Normal	735
Liver MVI	LiverMVI	MVI, Non-MVI	1,515
Gallbladder polyp	GBPolyp	Benign, Malignant	1,147
Pancreatic cancer[c]	LEPset[8]	Pancreatic cancer, Non-pancreatic cancer	3,500
Kidney tumor	Kidney2K	Benign, Malignant	2,579
Uterine fibroid[d]	UF1990[10]	Uterine fibroid, Non-uterine fibroid	1,990
Ovarian tumor[e]	MMOTU[5]	Chocolate cyst, Simple cyst, Serous cystadenoma, Normal ovary, Teratoma, Mucinous cystadenoma, Theca cell tumor, High grade serous cystadenocarcinoma	1,469
Fetal plane[f]	Fetus12K[48]	Fetal abdomen, Fetal brain, Fetal femur, Fetal thorax, Maternal cervix, Others	12,400
Breast lesion	BUID[49–51]	Benign, Malignant	232

Applications with publicly available datasets are underlined.
[a]https://zenodo.org/records/8231412.
[b]https://zenodo.org/records/7272660.
[c]https://zenodo.org/records/8041285.
[d]https://data.mendeley.com/datasets/552zbvzwrk/1.
[e]https://drive.google.com/drive/folders/1c5n0fVKrM9-SZE1kacTXPt1pt844iAs1.
[f]https://zenodo.org/records/3904280.

an AUC of 86.15%, confirming its consistent superior performance on fully independent external data.

The consistent superior performance of URFM across such a diverse range of ultrasound tasks involving different anatomical regions demonstrates its robustness, adaptability, and versatility. This versatility is particularly valuable in clinical settings where ultrasound imaging is used to diagnose a wide variety of conditions. The comprehensive evaluation across multiple organs and diagnostic challenges demonstrates that URFM is not only effective but also reliable for real-world applications.

In summary, these detailed comparisons highlight URFM's capability to generalize well across different clinical tasks, reinforcing its potential as a universal foundation model for ultrasound image analysis. The significant performance gains over existing methods indicate that URFM can provide more accurate and reliable diagnostics, which is crucial for improving patient outcomes in clinical practice.

URFM functions through representation-based masked image modeling

URFM employs representation-based MIM for pre-training. To evaluate its effectiveness against other SSL strategies, we conducted studies from two aspects: the first examines the superiority of MIM compared to other SSL approaches (Figure 4A), and the second assesses the advantages of representation-based MIM over pixel-based MIM (Figure 4B). For the first aspect, we compare various SSL frameworks, including MAE[36] (a representative of the MIM approach), SimCLR,[28] MoCo-v3,[31] DINO,[63] and our proposed URFM as a reference. SimCLR adopts ResNet-101[64] and other models adopt ViT-Base[43] architecture, with all models pre-trained on ImageNet.[62] Figure 4A illustrates the performance comparisons on the ten downstream tasks, and the mean average F1 of each model across all ten tasks is presented as a dashed line. Among these SSL approaches, MAE achieved statistically higher performance than the others, indicating it attained the best average performance across 10 tasks.

For the second aspect, we investigated various reconstruction targets within the MIM framework. We evaluated three distinct targets: pixel, CLIP representation,[41] and BiomedCLIP representation.[40] Both of the latter two belong to representation-based MIM, with the final one being the chosen target for URFM. All models use the ViT-Base architecture pre-trained on our ultrasound pre-training dataset. As shown in Figure 4B, BiomedCLIP representations consistently achieved higher performance metrics than the pixel-based approach across all tasks ($P < 0.05$). Furthermore, the mean average performance across the ten tasks indicates that BiomedCLIP is statistically superior to both CLIP and pixel-based approaches. These results indicate the superiority of representation-based MIM over pixel-based MIM, and are consistent with previous findings in non-medical domains.[65–67] In summary, these experiments substantiate the overall superiority of representation-based MIM, both in terms of outperforming other SSL approaches and enhancing the performance of pixel-based MIM.

URFM overcomes the "organ barrier"

Self-modality MIM pre-training[21,68,69] has proven effective for ultrasound imaging, i.e., pre-training ultrasound foundation models on ultrasound images demonstrate superiority over models pre-trained on nature images (ImageNet[62]). Since ultrasound imaging has been widely adopted across many organs covering the entire human body, a natural question arises: can a foundation model pre-trained on ultrasound images of one specific organ be effective for applications involving another organ? For example, can a thyroid foundation model generalize to the task of breast tumor diagnosis? With the introduction of URFM, we conducted a series of experiments to thoroughly investigate and address this question, and the results are presented in Figure 5. The dataset details regarding "organ barrier" is presented in Table 3.

In Figure 5A, we first pre-trained six organ-specific ultrasound foundation models using their respective organ ultrasound images from our pre-training dataset. These models were then fine-tuned on their corresponding downstream applications, a process we refer to as "organ-specific transfer". The "ImageNet transfer" represents the MAE[36] pre-trained on ImageNet,[62] serving as a baseline for comparison. As expected, all organ-specific transfers show improved performance compared to the ImageNet transfer.

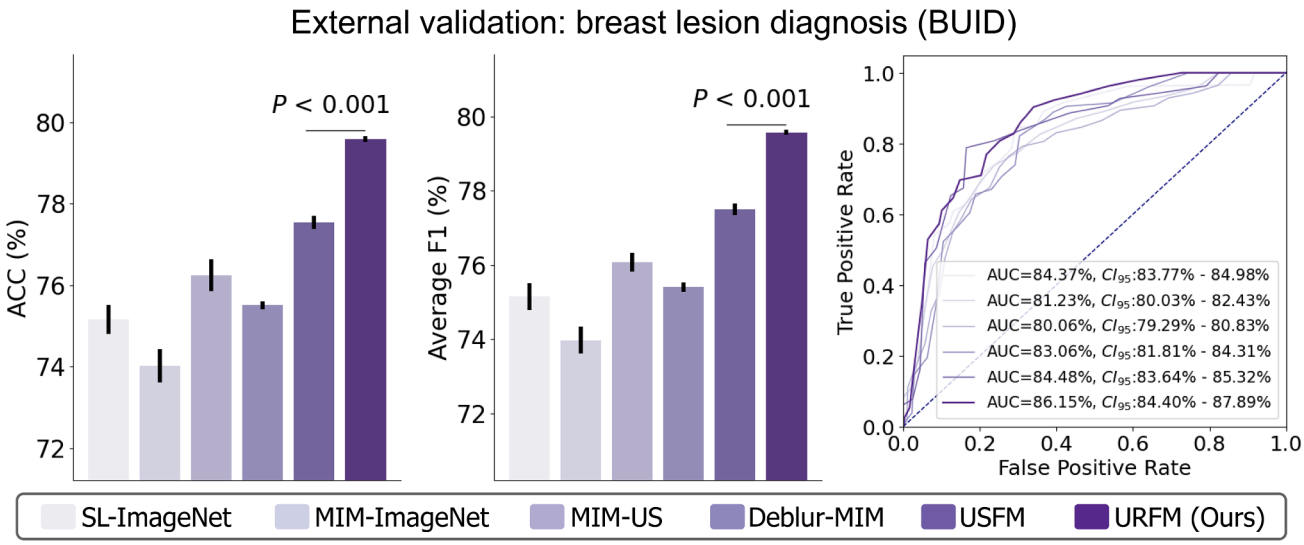


Figure 3. Experimental results of external validation on BUID dataset

Error bars represent the 95% CI. Statistical significance is assessed using *P* values calculated from two-sided *t*-tests.

Subsequently, we transferred these organ-specific foundation models to applications involving more organs, with the results shown in Figures 5B and 5C. Figure 5B illustrates the actual performance differences compared with the ImageNet transfer baseline, while Figure 5C shows the normalized performance differences, scaled to [-1, 1]. As demonstrated, most organ-specific foundation models perform poorly on tasks involving other organs beyond the pre-trained organ, with the only exceptions being the liver and gallbladder, where the two organs are anatomically close to each other. These findings indicate the existence of an "organ barrier" between pre-training and downstream transfer in MIM, i.e., organs not included in the pre-training data may perform poorly in downstream tasks. This emphasizes the necessity of constructing a comprehensive pre-training dataset that covers multiple organs, as we have done in our URFM.

Interpreting URFM through attention maps and feature embeddings

We provide two visualizations to intuitively demonstrate URFM's superiority. The first shows attention maps from both pre-training and downstream fine-tuning stages, as shown in Figure 6. In pre-training, these maps highlight encoded ultrasound representations from BiomedCLIP and reconstructed representations from URFM, while class activation maps (CAM)[70] are used for fine-tuning. Figure 6 reveals significant consistency between both stages, with attention focused primarily on lesion regions or key anatomical areas, essential for accurate diagnosis. This consistency highlights the effectiveness of URFM's representation-based MIM in enhancing downstream fine-tuning performance.

Furthermore, we also utilize t-SNE[71] to visualize feature embeddings at three key stages of URFM: before pre-training, after pre-training, and after fine-tuning (Figure 7). Comparing embeddings before and after pre-training reveals that images from the same organ begin to cluster, indicating the pre-training's success in aligning organ-specific features. After fine-tuning, these clusters further separate into clinically meaningful categories (e.g., benign vs. malignant), guided by downstream task labels. This progression in feature alignment and clustering underscores URFM's effectiveness in refining organ-specific representations and enhancing diagnostic capabilities post fine-tuning.

Pre-training scaling and downstream efficiency in URFM

The results of pre-training scaling and downstream efficiency of our URFM are demonstrated in Figure 8. The pre-training and downstream fine-tuning settings are presented in Tables 4 and 5, respectively. Pre-training scaling encompasses data scaling (Figure 8A), where different percentages of pre-training data are used, and model size scaling (Figure 8B), which involves different ViT architectures (ViT-Base, ViT-Large, and ViT-Huge) for pre-training. Downstream efficiency is evaluated by label efficiency (Figure 8C), measuring performance with varying percentages of fine-tuning data, and training time efficiency (Figure 8D), assessing convergence time during fine-tuning.

In the pre-training data scaling experiments (Figure 8A), we used 10%, 20%, 50%, and 100% of the pre-training data. Given the varying magnitudes of organ data within the pre-training dataset, we selected four representative downstream tasks for evaluation: ovarian tumor (3,593 images), pancreatic cancer (29,639 images), kidney tumor (111,209 images), and thyroid nodule (443,985 images). The number of pre-trained organ images corresponding to these tasks ranged from 3,593 to 443,985. The overall trend shows that performance improves with more pre-training data, aligning with established principles of data scaling in SSL.[72,73] Our proposed URFM model, utilizing 100% of the pre-training data, demonstrates statistically superior performance compared to other models. Notably, even with reduced pre-training data (50%, 20%, or 10%), URFM often outperforms others; for instance, in the thyroid nodule task, it achieves the highest performance using only 10% of the pre-training data. However, further analysis reveals nuanced

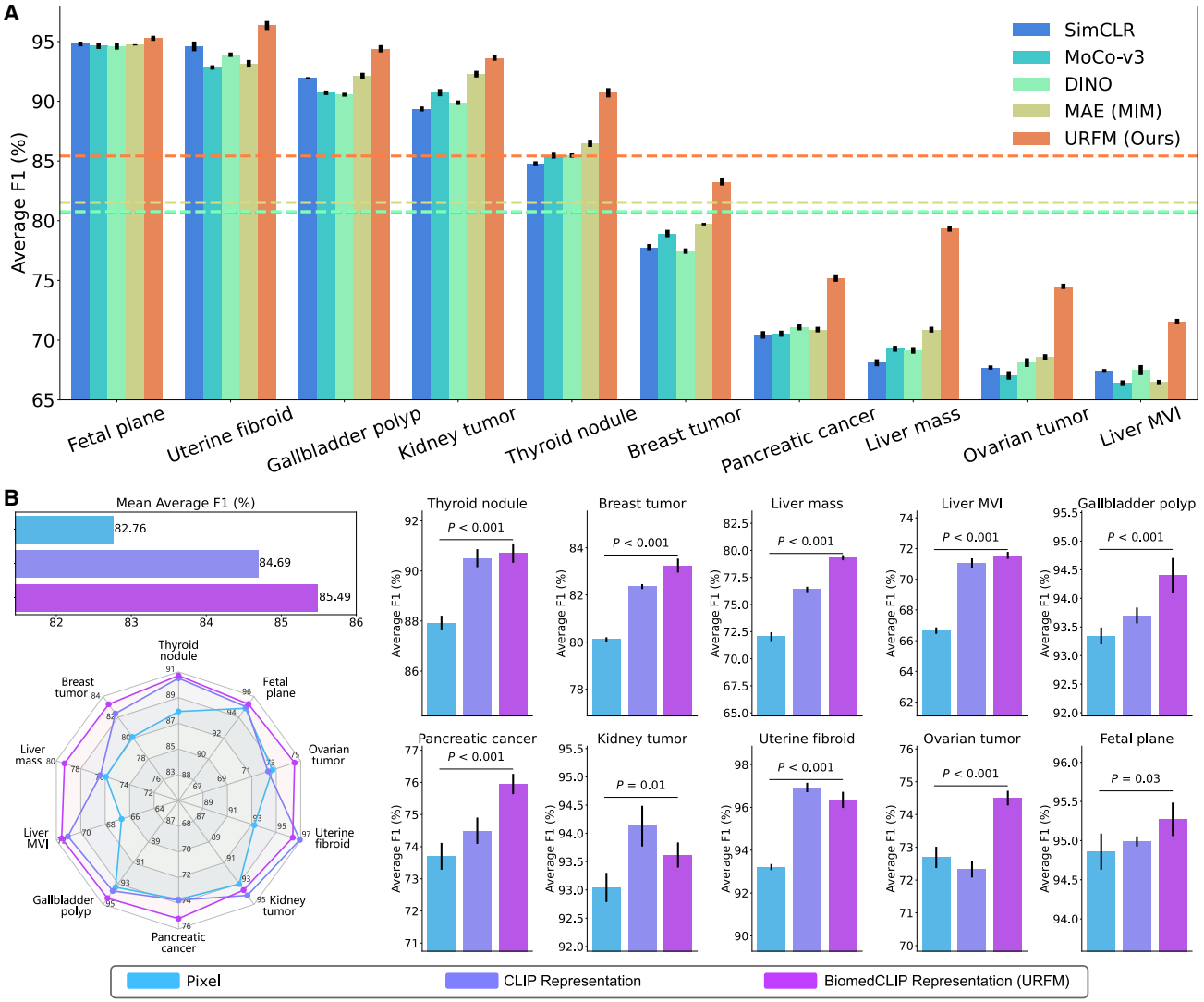


Figure 4. The ablations experiments of our representation-based MIM include the following

(A) The superiority of MIM compared to other SSL approaches. SimCLR adopts ResNet-101[64] and other models adopt ViT-Base[43] architecture, all models are pre-trained on ImageNet.[62]

(B) The superiority of representation-based MIM over pixel-based MIM. All models use the ViT-Base architecture pre-trained on our ultrasound pre-training dataset. All *P* values are calculated between the representation (BiomedCLIP) and pixel to check statistical significance. Error bars represent the 95% CI. Statistical significance is assessed using *P* values calculated from two-sided *t*-tests.

differences among the specific downstream tasks. For tasks pre-trained with less than 100,000 images (ovarian tumor and pancreatic cancer), performance gains are more significant. Conversely, tasks pre-trained with over 100,000 images (kidney tumor and thyroid nodule), show minimal improvement beyond certain performance thresholds. For example, performance gains for kidney tumor from 50% to 100% pre-training data, and for thyroid nodule from 20% to 100% are marginal. These findings, in line with previous MIM data scaling research for nature images,[72,74] suggest a data scaling limit in ultrasound imaging. Specifically, each organ in ultrasound exhibits a threshold beyond which additional pre-training data yields diminishing returns in performance improvement.

In the model size scaling experiments (Figure 8B), we evaluate performance by increasing the size of the pre-training model, using three ViT variants: ViT-Base, ViT-Large, and ViT-Huge.[43] Six downstream tasks of varying difficulty were selected: liver MVI, ovarian tumor, pancreatic cancer, liver mass, fetal plane, and uterine fibroid. The first four are considered as 'difficult' (Average F1 < 0.85), while the last two are "easy". The URFM with ViT-Huge architecture shows statistically significant improvements over the smaller models in the first four "difficult" tasks. However, in the last two "easy" tasks, URFM (ViT-Huge) does not outperform URFM (ViT-Large), suggesting that scaling model size benefits more in challenging tasks. The results in Figure 8B clearly illustrate that for the "difficult" tasks, performance

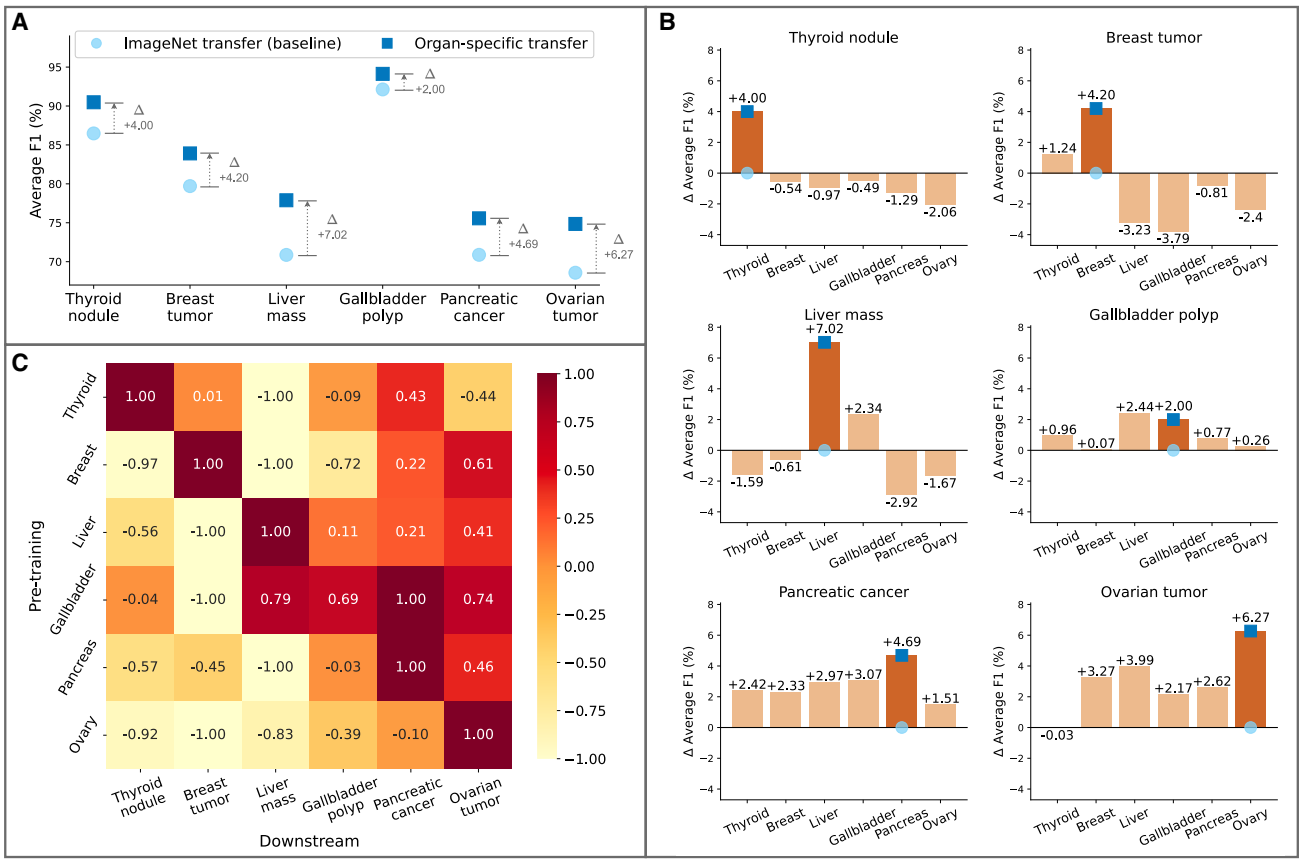


Figure 5. Experimental results of the "organ barrier"

(A) Comparison of organ-specific transfer versus ImageNet transfer: six organ-specific ultrasound foundation models were pre-trained on their respective organ ultrasound images from our pre-training dataset and fine-tuned on corresponding downstream tasks. All organ-specific transfers outperformed the "ImageNet transfer" baseline (the MAE[36] pre-trained on ImageNet[62]).

(B) Performance of organ-specific foundation models on applications involving different organs, showing significant declines when transferred to unrelated organs, except for anatomically close organs like the liver and gallbladder.

(C) Normalized performance differences, highlighting the "organ barrier" in transfer learning, underscoring the need for a comprehensive pre-training dataset covering multiple organs.

improves significantly with larger models. However, for the "easy" tasks, the performance gains from ViT-Large to ViT-Huge are minimal, with ViT-Huge, even slightly underperforming ViT-Large on the fetal plane task. These findings suggest that model size scaling benefits more challenging tasks, while easier tasks may reach a performance plateau, rendering further model size increases less impactful.

Downstream efficiency is evaluated in two aspects: label efficiency and training time efficiency. Label efficiency (Figure 8C) is analyzed by comparing model performance across varying percentages of labeled fine-tuning data. URFM consistently exhibits statistically better performance across different label percentages compared to other models. We tested four tasks, ranging from liver mass (735 images) to fetal plane (12,400 images). URFM consistently delivers superior performance across different labeled data percentages, indicating its strong adaptability with minimal annotation requirements. Training time efficiency (Figure 8D) evaluates convergence speed during fine-tuning. URFM achieves statistically faster convergence speeds relative to other models, indicating enhanced training efficiency. These results underscore URFM's exceptional efficiency in both label and training time requirements.

DISCUSSION

In this study, we introduce URFM, a general ultrasound foundation model pre-trained using representation-based MIM on a large-scale, diverse ultrasound dataset. URFM significantly

Table 3. Summary of pre-training data for the "organ barrier" experiment

Organ	Thyroid	Breast	Liver	Gallbladder	Pancreas	Ovary
Images	89,985	84,663	74,852	39,743	29,639	3,593

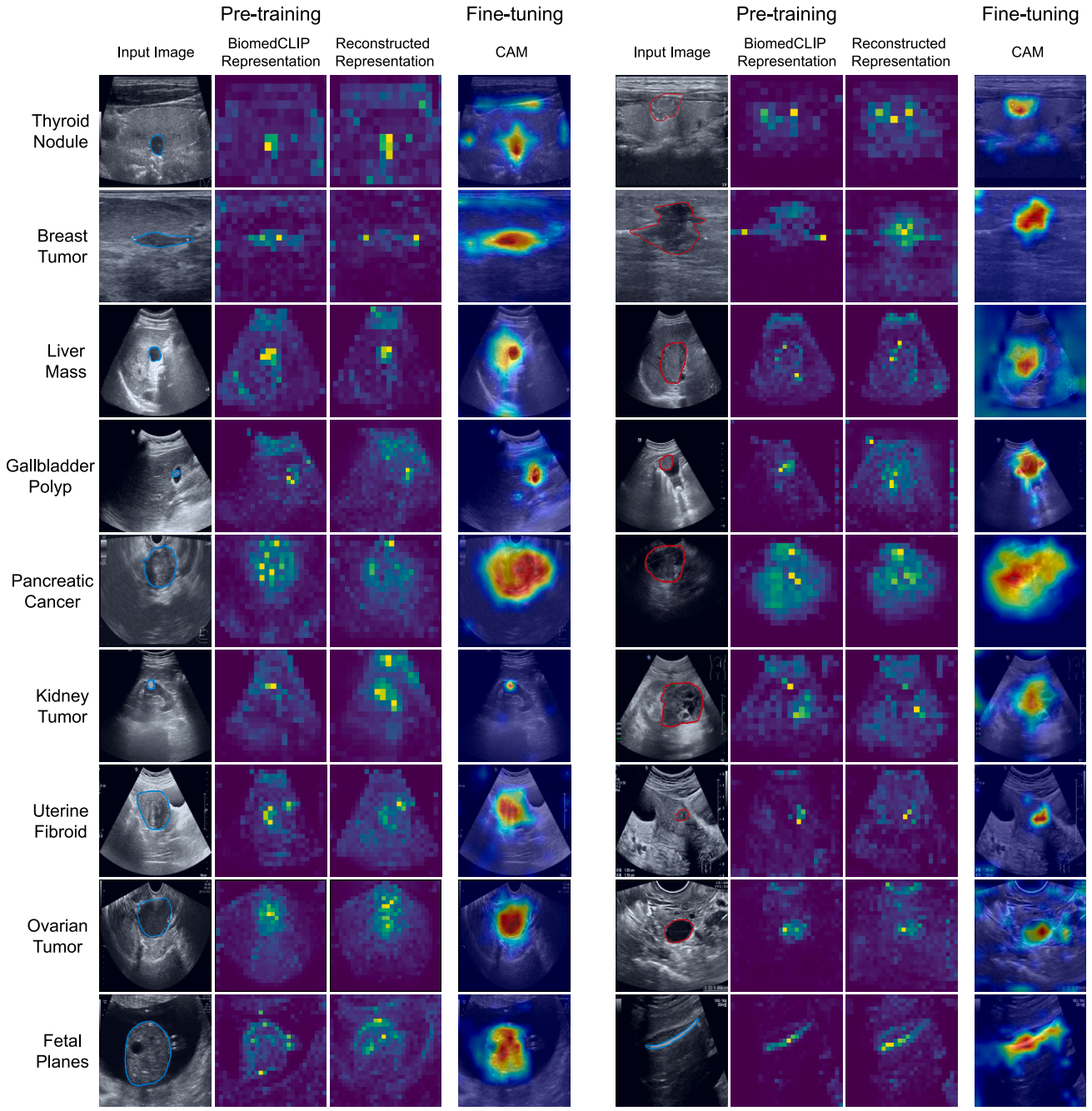


Figure 6. Attention maps from both pre-training and downstream fine-tuning stages of URFM

The colored lines in the input image denote the boundaries of each lesion or anatomical organ, with blue representing benign lesions and red representing malignant lesions. For the attention maps of pre-training, we visualize two representations: BiomedCLIP representations, which serve as the reconstruction targets during URFM pre-training, and the reconstructed representations by URFM. These representations are visualized through the self-attention of the [CLS] token in the heads of the last layer of the corresponding ViT models. For the attention maps of fine-tuning, we use the widely adopted class activation map (CAM)[70] of the last layer of the corresponding fine-tuned ViT model for each task.

advances ultrasound image analysis, achieving notable improvements in diagnostic accuracy across a wide range of clinical applications involving nine different organs. By leveraging representation-based MIM and extensive, varied ultrasound data for pre-training, URFM effectively captures general and meaningful ultrasound representations, outperforming other foundation models during downstream fine-tuning. Its exceptional performance across ten distinct tasks highlights its robustness and versatility, establishing URFM as a new foundation for effective and accurate ultrasound image analysis.

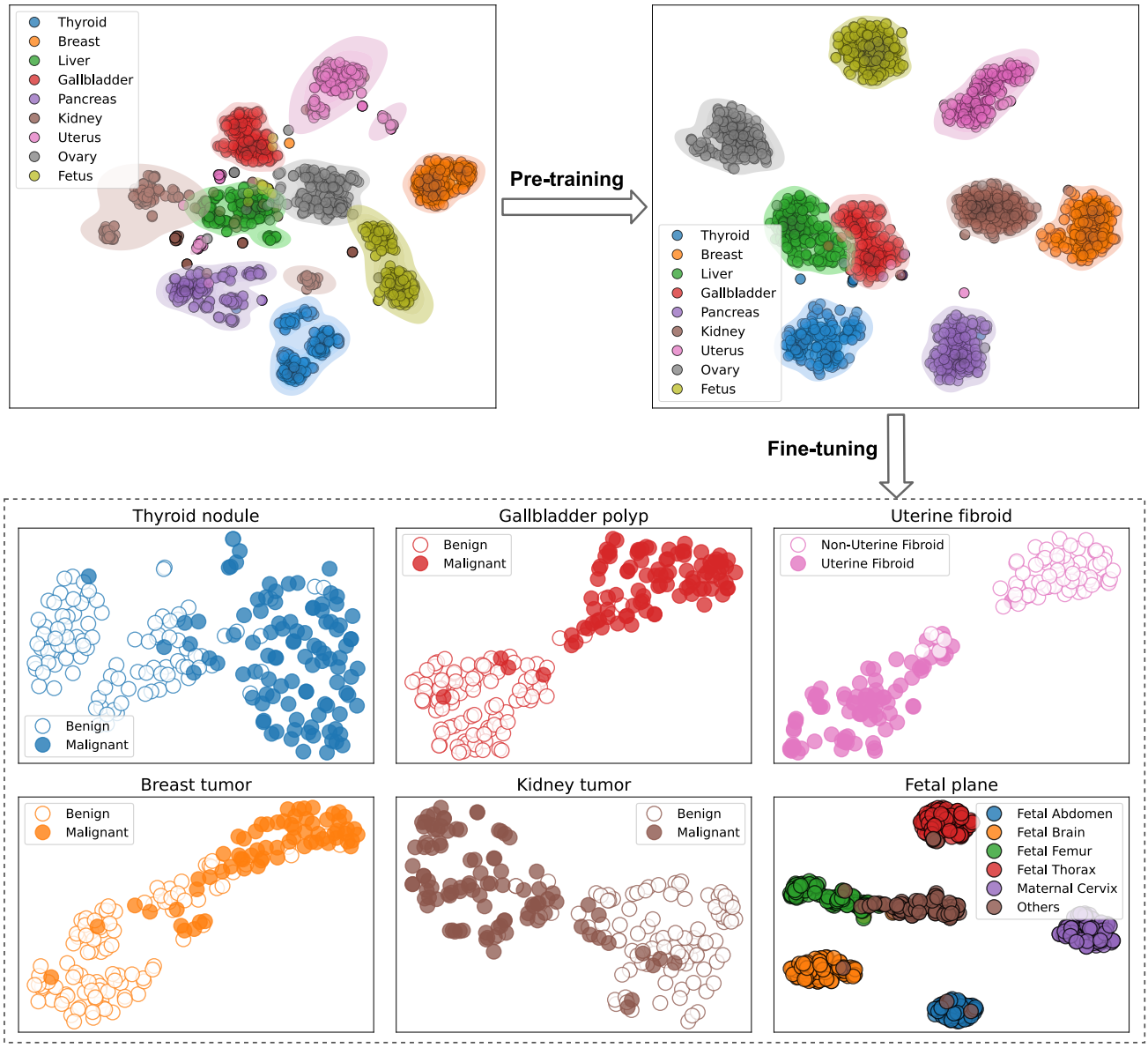


Figure 7. t-SNE visualization of the feature embeddings at three key stages of URFM: before pre-training, after pre-training, and after fine-tuning

The MAE[36] pre-trained on ImageNet is used as the model for the before pre-training stage. For the first two stages, 200 images are randomly selected from nine downstream datasets: Thyroid4K,[4] BUS-BRA,[47] Liver735,[7] GBPolyp, LEPset,[8] Kidney2K, UF1990,[10] MMOTU,[5] and Fetus12K,[48] with each dataset representing a different organ. For the after fine-tuning stage, images for computing embeddings are randomly selected from the validation set of each corresponding downstream dataset.

A key innovation of URFM is its adoption of representation-based MIM, which offers substantial improvements over conventional pixel-based MIM approach, enhancing both the efficiency and effectiveness of feature extraction and representation in ultrasound imaging. Moreover, the superior ultrasound feature representation achieved with BiomedCLIP,[40] compared to general CLIP,[41] can be attributed to its specialized biomedical knowledge. This expertise enriches the model's capability to understand, interpret and extract complex image patterns in ultrasound, thereby providing meaningful and valuable representations as the reconstruction target for the URFM pre-training.

In addition, the study identifies a notable challenge known as the "organ barrier" in ultrasound imaging, where models pre-trained on images from specific organs exhibit reduced performance on tasks involving different organs. URFM mitigates this issue by using a diverse range of anatomical organs in its pre-training data, thereby overcoming the limitations of organ-specific models and ensuring robust generalization performance across a broad spectrum of diagnostic tasks. This cross-organ

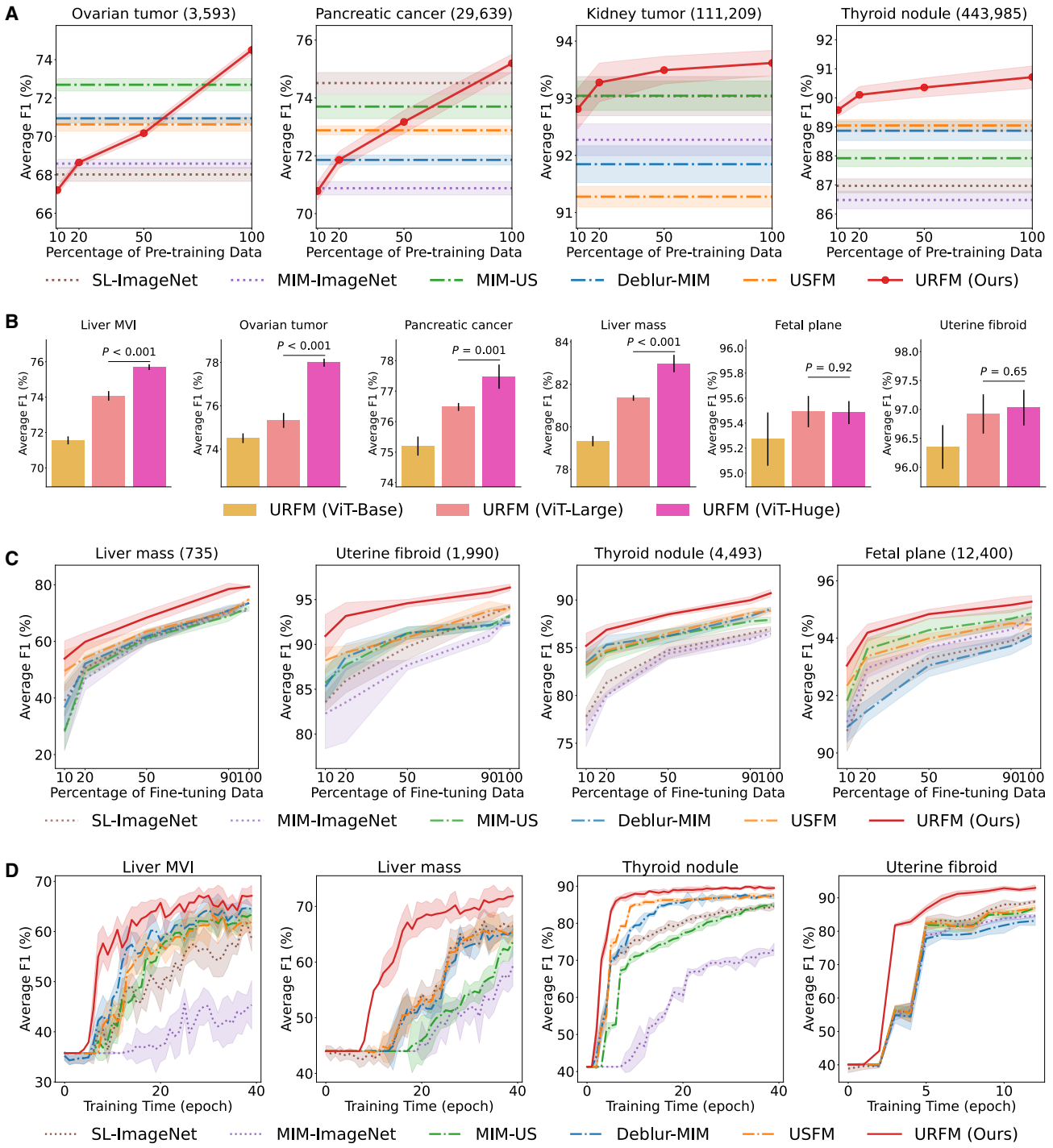


Figure 8. Pre-training scaling and downstream efficiency of our URFM

(A) Pre-training data scaling, which employs different percentages of the pre-training data. The numbers after each task name indicate the corresponding organ images in the pre-training dataset.

(B) Pre-training model size scaling, which adopts different ViT architectures (ViT-Base, ViT-Large, and ViT-Huge) for pre-training.

(C) Downstream label efficiency, which measures performance across varying percentages of downstream fine-tuning data. The numbers next to each task name represent the total images for this downstream task.

(D) Downstream training time efficiency, which assesses the training time required for convergence during fine-tuning. Shaded areas and error bars represent the 95% CI.

Table 4. URFM pre-training settings

Config	Value
Masking ratio	75%
Input size	224 × 224
Optimizer	AdamW[52]
Base learning rate	1.5e–4
Weight decay	0.05
Optimizer momentum	β_1, β_2 = 0.9, 0.95[53]
Batch size	1024
Learning rate schedule	cosine decay[54]
Layer-wise learning rate decay[55]	1.0
Warmup epochs[56]	40
Total pre-training epochs	400
Augmentation	RandomResizedCrop

URFM uses ViT-B/16[43] as the encoder by default.

Table 5. URFM downstream fine-tuning settings

Config	Value
Optimizer	AdamW[52]
Base learning rate	5e–4
Weight decay	0.05
Optimizer momentum	β_1, β_2 = 0.9, 0.999
Batch size	16
Learning rate schedule	cosine decay[54]
Layer-wise learning rate decay[55]	0.75
Warmup epochs	5
Augmentation	RandAug (9, 0.5)[57]
Label smoothing[58]	0.1
Mixup[59]	0.8
Cutmix[60]	1.0
Drop path[61]	0.1

We use full fine-tuning for downstream transfer.

generalization capability is further enhanced by URFM's representation-based pre-training using BiomedCLIP, which enables the model to learn domain-specific semantic representations instead of relying on raw noisy pixels. Our experiments demonstrate that this approach leads to superior downstream performance compared to both traditional MIM and general-purpose vision-language models like CLIP. This adaptability is crucial for real-world clinical applications, where foundation models must perform effectively across diverse organ types and conditions.

Further, we address the issue of underrepresentation of certain organs, such as veins, in our dataset. While we explored strategies like weighted sampling to balance the data, these methods did not result in significant improvements in performance, suggesting that URFM's representation-based pre-training is robust to such imbalances. This indicates that the model is capable of handling underrepresented organ categories effectively, though exploring advanced balancing techniques remains a promising direction for future work.

The study also highlights the influence of data scaling and model size scaling in pre-training on URFM's downstream performance. Consistent with SSL principles, the model exhibits enhanced downstream performance as pre-training data volume increases. However, for each specific organ in ultrasound, the benefits begin to diminish at higher organ-specific data sizes, suggesting an optimal balance between organ-specific data and pre-training efficiency. This highlights the importance of strategic dataset curation to maximize performance. Additionally, scaling model size demonstrates URFM's adaptability: larger models excel in complex tasks, while gains for simpler tasks plateau, emphasizing the need to match model size with task complexity.

URFM also demonstrates exceptional efficiency in downstream fine-tuning, both in label and time efficiency. The model consistently performs well across tasks with varying amounts of labeled data, highlighting its strong label efficiency—particularly valuable in scenarios with limited annotated data. Additionally, URFM's rapid convergence in fine-tuning significantly reduces computational costs, making it well-suited for practical clinical applications where efficient use of data and resources is essential.

Further analysis of attention maps and feature embeddings reveals insights into URFM's extraordinary performance. The consistency of attention maps between pre-training and fine-tuning indicates that URFM effectively learns and retains critical image features and patterns, enhancing its downstream performance. In addition, the t-SNE visualizations illustrate the model's ability to cluster organ-specific features and refine these clusters during fine-tuning.

In conclusion, URFM represents a significant advancement in ultrasound image analysis by utilizing an advanced SSL approach and extensive pre-training to achieve exceptional performance across various clinical applications. The model's generalizability, versatility, and efficiency highlight its potential to improve diagnostic accuracy and efficiency in ultrasound-related clinical scenarios. We anticipate that the foundation model, URFM, will lead to improved clinical outcomes and more effective utilization of ultrasound imaging in the medical field.

Limitations of the study

Despite its considerable success in ultrasound diagnostics, URFM exhibits certain limitations that warrant further investigation. In particular, its extension to fine-grained tasks —such as segmentation and detection, which require precise localization and detailed feature extraction— remains challenging. Moreover, the model's performance on downstream tasks involving organs not included in the pre-training dataset has yet to be fully explored. Additionally, although increasing the dataset size and model capacity generally enhances performance, the benefits tend to plateau for specific organ tasks and simpler applications. These observations suggest that beyond a certain threshold, merely scaling up may not yield proportional gains, emphasizing the need for more efficient and tailored pre-training strategies.

RESOURCE AVAILABILITY

Lead contact

Requests for further information and resources should be directed to and will be fulfilled by the lead contact, Kang Li (likang@wchscu.cn).

Materials availability

This study did not generate new materials.

Data and code availability

- The data used for the pre-training and downstream fine-tuning are provided in Tables 1 and 2. Most of the data are publicly available, except for the private data. All private data reported in this paper will be shared by the lead contact upon request.
- All original code and pre-trained models are publicly accessible on Github.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

ACKNOWLEDGMENTS

This work was supported by the National Key Research and Development Program of China under Grant 2020YFB1711500, the 1 · 3 · 5 project for disciplines of excellence, West China Hospital, Sichuan University(ZYYC21004), and the Natural Science Foundation of Sichuan Province under Grant NO. 2023NSFSC1722.

AUTHOR CONTRIBUTIONS

Q. Kang, Q. Lao, and K.L. conceived the project; Q. Kang conducted the model training and evaluation experiments, and, together with C.D. and Z.H., carried out the validation studies. Q. Kang, Q. Lao, and J.G. analyzed the results. Q.L. and W.B. curated and annotated the data. J.G., W.B., and C.D. contributed to the interpretation and visualization. Q. Kang wrote the manuscript assisted by J.G., C.D., and Z.H. under the direction of Q. Lao, K.L., and Q.L. All the authors contributed to this work through useful discussion and/or comments on the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- METHOD DETAILS
  - Datasets
  - Pre-training of URFM
  - Downstream fine-tuning of URFM
  - Performance evaluation
- QUANTIFICATION AND STATISTICAL ANALYSIS

Received: January 7, 2025
Revised: April 24, 2025
Accepted: June 13, 2025
Published: June 18, 2025

REFERENCES

1. Jensen, J.A. (2007). Medical ultrasound imaging. Prog. Biophys. Mol. Biol. *93*, 153–165.
2. Carovac, A., Smajlovic, F., and Junuzovic, D. (2011). Application of ultrasound in medicine. Acta Inform. Med. *19*, 168–171.
3. Akkus, Z., Cai, J., Boonrod, A., Zeinoddini, A., Weston, A.D., Philbrick, K. A., and Erickson, B.J. (2019). A survey of deep-learning applications in ultrasound: Artificial intelligence–powered ultrasound for improving clinical workflow. J. Am. Coll. Radiol. *16*, 1318–1328.
4. Kang, Q., Lao, Q., Li, Y., Jiang, Z., Qiu, Y., Zhang, S., and Li, K. (2022). Thyroid nodule segmentation and classification in ultrasound images through intra-and inter-task consistent learning. Med. Image Anal. *79*, 102443.
5. Zhao, Q., Lyu, S., Bai, W., Cai, L., Liu, B., Wu, M., Sang, X., Yang, M., and Chen, L. (2022). A multi-modality ovarian tumor ultrasound image dataset for unsupervised cross-domain semantic segmentation. Preprint at arXiv. https://doi.org/10.48550/arXiv.2207.06799.
6. Gao, J., Lao, Q., Kang, Q., Liu, P., Zhang, L., and Li, K. (2022). Unsupervised cross-disease domain adaptation by lesion scale matching. In International Conference on Medical Image Computing and Computer-Assisted Intervention (Springer), pp. 660–670.
7. Xu, Y., Zheng, B., Liu, X., Wu, T., Ju, J., Wang, S., Lian, Y., Zhang, H., Liang, T., Sang, Y., et al. (2023). Improving artificial intelligence pipeline for liver malignancy diagnosis using ultrasound images and video frames. Brief. Bioinform. *24*, bbac569.
8. Li, J., Zhang, P., Wang, T., Zhu, L., Liu, R., Yang, X., Wang, K., Shen, D., and Sheng, B. (2023). Dsmt-net: Dual self-supervised multi-operator transformation for multi-source endoscopic ultrasound diagnosis. IEEE Trans. Med. Imag. *43*, 64–75.
9. Gao, J., Lao, Q., Liu, P., Yi, H., Kang, Q., Jiang, Z., Wu, X., Li, K., Chen, Y., and Zhang, L. (2023). Anatomically guided cross-domain repair and screening for ultrasound fetal biometry. IEEE J. Biomed. Health Inform. *27*, 4914–4925.
10. Cai, P., Yang, T., Xie, Q., Liu, P., and Li, P. (2024). A lightweight hybrid model for the automatic recognition of uterine fibroid ultrasound images based on deep learning. J. Clin. Ultrasound *52*, 753–762.
11. Rajpurkar, P., Chen, E., Banerjee, O., and Topol, E.J. (2022). Ai in health and medicine. Nat. Med. *28*, 31–38.
12. van Sloun, R.J.G., Cohen, R., and Eldar, Y.C. (2020). Deep learning in ultrasound imaging. Proc. IEEE *108*, 11–29.
13. Azad, B., Azad, R., Eskandari, S., Bozorgpour, A., Kazerouni, A., Rekik, I., and Merhof, D. (2023). Foundational models in medical imaging: A comprehensive survey and future vision. Preprint at arXiv. https://doi.org/10.48550/arXiv.2310.18689.
14. Moor, M., Banerjee, O., Abad, Z.S.H., Krumholz, H.M., Leskovec, J., Topol, E.J., and Rajpurkar, P. (2023). Foundation models for generalist medical artificial intelligence. Nature *616*, 259–265.
15. Zhang, S., and Metaxas, D. (2024). On the challenges and perspectives of foundation models for medical image analysis. Med. Image Anal. *91*, 102996.
16. Wang, X., Zhang, X., Wang, G., He, J., Li, Z., Zhu, W., Guo, Y., Dou, Q., Li, X., Wang, D., et al. (2024). Openmedlab: An open-source platform for multi-modality foundation models in medicine. Preprint at arXiv. https://doi.org/10.48550/arXiv.2402.18028.
17. 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., et al. (2023). A foundation model for generalizable disease detection from retinal images. Nature *622*, 156–163.
18. Vorontsov, E., Bozkurt, A., Casson, A., Shaikovski, G., Zelechowski, M., Severson, K., Zimmermann, E., Hall, J., Tenenholtz, N., Fusi, N., et al. (2024). A foundation model for clinical-grade computational pathology and rare cancers detection. Nat. Med. *30*, 2924–2935.
19. Chen, R.J., Ding, T., Lu, M.Y., Williamson, D.F.K., Jaume, G., Song, A.H., Chen, B., Zhang, A., Shao, D., Shaban, M., et al. (2024). Towards a general-purpose foundation model for computational pathology. Nat. Med. *30*, 850–862.
20. Pai, S., Bontempi, D., Hadzic, I., Prudente, V., Sokač, M., Chaunzwa, T.L., Bernatz, S., Hosny, A., Mak, R.H., Birkbak, N.J., and Aerts, H.J.W.L. (2024). Foundation model for cancer imaging biomarkers. Nat. Mach. Intell. *6*, 354–367.
21. Jiao, J., Zhou, J., Li, X., Xia, M., Huang, Y., Huang, L., Wang, N., Zhang, X., Zhou, S., Wang, Y., and Guo, Y. (2024). Usfm: A universal ultrasound

foundation model generalized to tasks and organs towards label efficient image analysis. Med. Image Anal. *96*, 103202.

22. Ma, C., Tan, W., He, R., and Yan, B. (2024). Pretraining a foundation model for generalizable fluorescence microscopy-based image restoration. Nat. Methods *21*, 1558–1567.
23. Xu, H., Usuyama, N., Bagga, J., Zhang, S., Rao, R., Naumann, T., Wong, C., Gero, Z., González, J., Gu, Y., et al. (2024). A whole-slide foundation model for digital pathology from real-world data. Nature *630*, 181–188.
24. Yao, J., Wang, X., Song, Y., Zhao, H., Ma, J., Chen, Y., Liu, W., and Wang, B. (2024). Eva-x: A foundation model for general chest x-ray analysis with self-supervised learning. Preprint at arXiv. https://doi.org/10.48550/arXiv.2405.05237.
25. Liu, X., Zhang, F., Hou, Z., Mian, L., Wang, Z., Zhang, J., and Tang, J. (2021). Self-supervised learning: Generative or contrastive. IEEE Trans. Knowl. Data Eng. *35*, 857–876.
26. Gui, J., Chen, T., Zhang, J., Cao, Q., Sun, Z., Luo, H., and Tao, D. (2024). A survey on self-supervised learning: Algorithms, applications, and future trends. IEEE Trans. Pattern Anal. Mach. Intell. *46*, 9052–9071.
27. Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. (2020). A simple framework for contrastive learning of visual representations. In International conference on machine learning, pp. 1597–1607.
28. Chen, T., Kornblith, S., Swersky, K., Norouzi, M., and Hinton, G.E. (2020). Big self-supervised models are strong semi-supervised learners. Adv. Neural Inf. Process. Syst. *33*, 22243–22255.
29. He, K., Fan, H., Wu, Y., Xie, S., and Girshick, R. (2020). Momentum contrast for unsupervised visual representation learning. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 9729–9738.
30. Chen, X., Fan, H., Girshick, R., and He, K. (2020). Improved baselines with momentum contrastive learning. Preprint at arXiv. https://doi.org/10.48550/arXiv.2003.04297.
31. Chen, X., Xie, S., and He, K. (2021). An empirical study of training self-supervised vision transformers. In Proceedings of the IEEE/CVF international conference on computer vision, pp. 9640–9649.
32. Li, Y., Lao, Q., Kang, Q., Jiang, Z., Du, S., Zhang, S., and Li, K. (2023). Self-supervised anomaly detection, staging and segmentation for retinal images. Med. Image Anal. *87*, 102805.
33. Li, Y., Qian, G., Jiang, X., Jiang, Z., Wen, W., Zhang, S., Li, K., and Lao, Q. (2024). Hierarchical-instance contrastive learning for minority detection on imbalanced medical datasets. IEEE Trans. Med. Imaging *43*, 416–426.
34. Bao, H., Dong, L., Piao, S., and Wei, F. (2021). Beit: Bert pre-training of image transformers. Preprint at arXiv. https://doi.org/10.48550/arXiv.2106.08254.
35. Xie, Z., Zhang, Z., Cao, Y., Lin, Y., Bao, J., Yao, Z., Dai, Q., and Hu, H. (2022). Simmim: A simple framework for masked image modeling. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 9653–9663.
36. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., and Girshick, R. (2022). Masked autoencoders are scalable vision learners. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 16000–16009.
37. Kang, Q., Gao, J., Li, K., and Lao, Q. (2023). Deblurring masked autoencoder is better recipe for ultrasound image recognition. In International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 352–362.
38. Kang, Q., Lao, Q., Gao, J., Liu, J., Yi, H., Ma, B., Zhang, X., and Li, K. (2024). Deblurring masked image modeling for ultrasound image analysis. Med. Image Anal. *97*, 103256.
39. Gao, P., Ma, T., Li, H., Dai, J., and Qiao, Y. (2022). Convmae: Masked convolution meets masked autoencoders. Preprint at arXiv. https://doi.org/10.48550/arXiv.2205.03892.
40. Zhang, S., Xu, Y., Usuyama, N., Xu, H., Bagga, J., Tinn, R., Preston, S., Rao, R., Wei, M., Valluri, N., et al. (2023). Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. Preprint at arXiv. https://doi.org/10.48550/arXiv.2303.00915.
41. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. (2021). Learning transferable visual models from natural language supervision. In International conference on machine learning, pp. 8748–8763.
42. Wang, L., Zhang, L., Zhu, M., Qi, X., and Yi, Z. (2020). Automatic diagnosis for thyroid nodules in ultrasound images by deep neural networks. Med. Image Anal. *61*, 101665.
43. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. Preprint at arXiv. https://doi.org/10.48550/arXiv.2010.11929.
44. Mei, X., Liu, Z., Robson, P.M., Marinelli, B., Huang, M., Doshi, A., Jacobi, A., Cao, C., Link, K.E., Yang, T., et al. (2022). Radimagenet: an open radiologic deep learning research dataset for effective transfer learning. Radiol. Artif. Intell. *4*, e210315.
45. Lin, Z., Lin, J., Zhu, L., Fu, H., Qin, J., and Wang, L. (2022). A new dataset and a baseline model for breast lesion detection in ultrasound videos. In International Conference on Medical Image Computing and Computer-Assisted Intervention (Springer), pp. 614–623.
46. Leclerc, S., Smistad, E., Pedrosa, J., Østvik, A., Cervenansky, F., Espinosa, F., Espeland, T., Berg, E.A.R., Jodoin, P.M., Grenier, T., et al. (2019). Deep learning for segmentation using an open large-scale dataset in 2d echocardiography. IEEE Trans. Med. Imaging *38*, 2198–2210.
47. Gómez-Flores, W., Gregorio-Calas, M.J., and Coelho de Albuquerque Pereira, W. (2023). Bus-bra: A breast ultrasound dataset for assessing computer-aided diagnosis systems. Med. Phys. *51*, 3110–3123.
48. Burgos-Artizzu, X.P., Coronado-Gutiérrez, D., Valenzuela-Alcaraz, B., Bonet-Carne, E., Eixarch, E., Crispi, F., and Gratacós, E. (2020). Evaluation of deep convolutional neural networks for automatic classification of common maternal fetal ultrasound planes. Sci. Rep. *10*, 10200.
49. Abbasian Ardakani, A., Mohammadi, A., Mirza-Aghazadeh-Attari, M., and Acharya, U.R. (2023). An open-access breast lesion ultrasound image database: Applicable in artificial intelligence studies. Comput. Biol. Med. *152*, 106438.
50. Hamyoon, H., Yee Chan, W., Mohammadi, A., Yusuf Kuzan, T., Mirza-Aghazadeh-Attari, M., Leong, W.L., Murzoglu Altintoprak, K., Vijayananthan, A., Rahmat, K., Ab Mumin, N., et al. (2022). Artificial intelligence, bi-rads evaluation and morphometry: A novel combination to diagnose breast cancer using ultrasonography, results from multi-center cohorts. Eur. J. Radiol. *157*, 110591.
51. Homayoun, H., Chan, W.Y., Kuzan, T.Y., Leong, W.L., Altintoprak, K.M., Mohammadi, A., Vijayananthan, A., Rahmat, K., Leong, S.S., Mirza-Aghazadeh-Attari, M., et al. (2022). Applications of machine-learning algorithms for prediction of benign and malignant breast lesions using ultrasound radiomics signatures: A multi-center study. Biocybern. Biomed. Eng. *42*, 921–933.
52. Loshchilov, I., and Hutter, F. (2017). Decoupled weight decay regularization. Preprint at arXiv. https://doi.org/10.48550/arXiv.1711.05101.
53. Chen, M., Radford, A., Child, R., Wu, J., Jun, H., Luan, D., and Sutskever, I. (2020). Generative pretraining from pixels. In International conference on machine learning, pp. 1691–1703.
54. Loshchilov, I., and Hutter, F. (2016). Sgdr: Stochastic gradient descent with warm restarts. Preprint at arXiv. https://doi.org/10.48550/arXiv.1608.03983.
55. Clark, K., Luong, M.T., Le, Q.V., and Manning, C.D. (2020). Electra: Pre-training text encoders as discriminators rather than generators. Preprint at arXiv. https://doi.org/10.48550/arXiv.2003.10555.
56. Goyal, P., Dollár, P., Girshick, R., Noordhuis, P., Wesolowski, L., Kyrola, A., Tulloch, A., Jia, Y., and He, K. (2017). Accurate, large minibatch sgd: Training imagenet in 1 hour. Preprint at arXiv. https://doi.org/10.48550/arXiv.1706.02677.

57. Cubuk, E.D., Zoph, B., Shlens, J., and Le, Q.V. (2020). Randaugment: Practical automated data augmentation with a reduced search space. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops, pp. 702–703.

58. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2818–2826.

59. Zhang, H., Cisse, M., Dauphin, Y.N., and Lopez-Paz, D. (2017). mixup: Beyond empirical risk minimization. Preprint at arXiv. https://doi.org/10.48550/arXiv.1710.09412.

60. Yun, S., Han, D., Oh, S.J., Chun, S., Choe, J., and Yoo, Y. (2019). Cutmix: Regularization strategy to train strong classifiers with localizable features. In Proceedings of the IEEE/CVF international conference on computer vision, pp. 6023–6032.

61. Huang, G., Sun, Y., Liu, Z., Sedra, D., and Weinberger, K.Q. (2016). Deep networks with stochastic depth. In Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14, pp. 646–661.

62. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In 2009 IEEE conference on computer vision and pattern recognition, pp. 248–255.

63. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., and Joulin, A. (2021). Emerging properties in self-supervised vision transformers. In Proceedings of the IEEE/CVF international conference on computer vision, pp. 9650–9660.

64. He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778.

65. Wei, L., Xie, L., Zhou, W., Li, H., and Tian, Q. (2022). Mvp: Multimodality-guided visual pre-training. In European conference on computer vision, pp. 337–353.

66. Hou, Z., Sun, F., Chen, Y.K., Xie, Y., and Kung, S.Y. (2022). Milan: Masked image pretraining on language assisted representation. Preprint at arXiv. https://doi.org/10.48550/arXiv.2208.06049.

67. Liu, X., Zhou, J., Kong, T., Lin, X., and Ji, R. (2022). Exploring target representations for masked autoencoders. Preprint at arXiv. https://doi.org/10.48550/arXiv.2209.03917.

68. Basu, S., Gupta, M., Madan, C., Gupta, P., and Arora, C. (2024). Focus-mae: Gallbladder cancer detection from ultrasound videos with focused masked autoencoders. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11715–11725.

69. Rahman, A., and Patel, V.M. (2024). Ultramae: Multi-modal masked autoencoder for ultrasound pre-training. In Medical Imaging with Deep Learning.

70. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE international conference on computer vision, pp. 618–626.

71. Van der Maaten, L., and Hinton, G. (2008). Visualizing data using t-sne. J. Mach. Learn. Res. *9*, 2579–2605.

72. Xie, Z., Zhang, Z., Cao, Y., Lin, Y., Wei, Y., Dai, Q., and Hu, H. (2023). On data scaling in masked image modeling. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10365–10374.

73. Fang, Y., Wang, W., Xie, B., Sun, Q., Wu, L., Wang, X., Huang, T., Wang, X., and Cao, Y. (2023). Eva: Exploring the limits of masked visual representation learning at scale. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 19358–19369.

74. Lu, C.Z., Jin, X., Hou, Q., Liew, J.H., Cheng, M.M., and Feng, J. (2023). Delving deeper into data scaling in masked image modeling. Preprint at arXiv. https://doi.org/10.48550/arXiv.2305.15248.

75. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. Adv. Neural Inf. Process. Syst. *30*, 2.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
RadImageNet	RadImageNet website	https://doi.org/10.1148/ryai.210315
BUV	CVA-Net GitHub	https://doi.org/10.1007/978-3-031-16437-8_59
ultrasoundcases	ultrasoundcases.info	N/A
LEPset	Zenodo	https://doi.org/10.1109/TMI.2023.3289859
CAMUS	CREATIS CAMUS	https://doi.org/10.1109/TMI.2019.2900516
Software and algorithms		
Pytorch	N/A	pytorch.org
Pre-trained models and code	URFM GitHub repo	This manuscript
Other		
NVIDIA V100 GPU	Nvidia Corp., Santa Clara, California, USA	N/A

METHOD DETAILS

Datasets

Construction of the pre-training dataset

Our pre-training dataset is a comprehensive collection of 1,003,465 ultrasound images, encompassing 15 major anatomical organs or body parts. These images have been meticulously sourced from a variety of distinct origins, ensuring a diverse and robust dataset that accurately represents different ultrasound imaging conditions and anatomical variations. The detailed numbers for each organ and source are illustrated in Table 1. The private images, which form a significant part of the pre-training dataset, were acquired at West China Hospital with ethical approval. Although the organ distribution is not perfectly balanced, to the best of our knowledge, this dataset represents the most relatively balanced distribution across different organs globally. This extensive and varied dataset serves as a solid foundation for the pre-training of URFM, enabling it to learn and generalize effectively across various clinical scenarios.

Downstream fine-tuning applications

We have developed a benchmark evaluation consisting of ten unique downstream ultrasound clinical applications to thoroughly assess the performance of URFM. These applications encompass a diverse array of diagnostic targets across nine organs, each presenting varying levels of diagnostic complexity. The details of these clinical applications, including dataset, classes, and number of images, are provided in Table 2. Notably, six of these applications are publicly available and are highlighted in blue. Below is a description of each application.

(1) **Thyroid nodule**: The dataset for this application is Thyroid4K, proposed by Kang et al.[4] and was collected from West China Hospital, China. It comprises 4,493 thyroid ultrasound images from 4,493 patients (only one image was selected per patient). Among these, 2,576 images contain benign nodules, and 1,917 images contain malignant nodules. Malignant cases have corresponding fine needle aspiration (FNA) results, providing direct classification labels. The labels of benign cases were verified by senior radiologists during clinical diagnosis. All images were acquired using GE Logiq E9 ultrasound machine.

(2) **Breast tumor**: This application utilizes the BUS-BRA dataset, originally proposed by Gomez et al.[47] The images were collected from hospitals in Brazil. We selected only one image per patient, resulting in 1,064 images, with 722 images containing benign tumors and 342 images containing malignant tumors. The classification annotations of the tumors have all been confirmed by biopsy. The images were acquired using four different ultrasound machines: GE Logiq 5, GE Logiq 7, Toshiba Aplio 300, and GE U-Systems.

(3) **Liver mass**: This application uses the Liver735 dataset, proposed by Xu et al.,[7] and collected from three hospitals in China. It consisting of 735 images, including 435 images of malignant masses, 200 images of benign masses, and 100 images of normal livers. The labels were confirmed through biopsy, post-surgery pathology, or enhanced imaging. The images were acquired using over 20 different ultrasound machines, including models from Hitachi, Aloka, Philips, SonoScape, Mindray, Esaote, and others.

(4) **Liver MVI**: A private dataset named LiverMVI, collected from West China Hospital in China, is utilized for this application. It contains 1,515 images from 408 patients, with 731 images from 208 patients showing MVI and 784 images from 200 patients

without MVI (non-MVI). All annotations of these images were verified by pathological results. The images were acquired using two ultrasound machines: Philips IU 22 and Mindray Resona 7.

(5) **Gallbladder polyp**: This application uses a private dataset: GBPolyp, which was collected from West China Hospital in China. It consists of 1,147 images from 398 patients, with 546 images from 200 patients containing benign polyps and 601 images from 198 patients containing malignant polyps. The annotations for malignant cases were obtained from biopsy or surgery, while clinical follow-up results were used for benign cases. Multiple ultrasound machines, including those from GE, Philips, Mindray, Canon, and SuperSonic, were utilized to acquire these images.

(6) **Pancreatic cancer**: The LEPset dataset,[8] used for this application, was collected at Changhai Hospital, Shanghai, China. It includes 3,500 pathologically proven labeled endoscopic ultrasound (EUS) images from 420 patients. The dataset is composed of 280 patients with pancreatic cancer (1,680 images) and 140 patients with non-pancreatic cancer (1,820 images). The images were acquired using four different EUS machines: Pentax, Fujifilm, Aloka, and Olympus.

(7) **Kidney tumor**: A private dataset, Kidney2K, was collected from West China Hospital in China for this application. This dataset contains 2,579 images, including 1680 images with benign tumors and 899 images with malignant tumors. All malignant cases have corresponding pathological results, while all benign cases have been verified through clinical diagnosis. These images were acquired using A variety of ultrasound machines, including Philips, GE, Mindray, Hitachi and SonoScape.

(8) **Uterine fibroid**: This application adopts the UF1990 dataset,[10] which consists of 1990 images from 871 patients, collected from three hospitals in Fujian Province, China. Among these images, 875 are of uterine fibroid, while 1,115 are non-uterine fibroid images. All images were annotated by a professional sonographer.

(9) **Ovarian tumor**: The MMOTU dataset[5] used for this application was collected at Beijing Shijitan Hospital, China. It contains 1,469 ovarian ultrasound images from 294 patients, all acquired using the Mindray Resona8 ultrasound machine.

(10) **Fetal plane**: The Fetus12K dataset[48] was adopted for this application. It is a maternal-fetal ultrasound image dataset comprising 12,400 images from 1,792 patients, collected from two hospitals in Spain. All images were manually labeled with six of the most widely used maternal-fetal anatomical planes by a senior maternal-fetal specialist. Images were acquired using four distinct ultrasound machines: GE Voluson E6, GE Voluson S8, GE Voluson S10, and Aloka.

(11) **Breast lesion**: The BUID dataset[49–51] is an open-access collection of breast ultrasound images intended for AI research. It comprises 232 images, with 123 images depicting benign lesions and 109 images depicting malignant lesions, all confirmed by histopathology. Images were acquired using AirPlorer Ultimate ultrasound machine.

The dataset splitting protocol for the train, validation, and test subsets follows a ratio of 3:1:1 for all ten datasets. For datasets with multiple images from a single patient, the splitting is conducted at the patient level to ensure that images from the same patient do not appear in multiple subsets. In addition, we employ stratified random sampling to ensure that the distribution of each class is consistent across the train, validation, and test subsets.

Pre-training data summary in the 'organ barrier' experiment

To further interpret the results in Figure 5, we provide a summary of the data used to pre-train each organ-specific foundation model in Table 3. The number of images varies substantially across organs, from 3,593 for the ovary to 89,985 for the thyroid. Despite this large discrepancy in data volume, the performance degradation observed when transferring across different organs indicates that the 'organ barrier' is not merely attributable to insufficient pre-training data. In other words, even substantially increasing the amount of ultrasound data from a single organ does not enable the resulting foundation model to generalize effectively to other organs. These findings underscore the necessity of multi-organ pre-training, as implemented in URFM, to overcome the 'organ barrier' and ensure robust cross-organ generalization.

Pre-training of URFM

The pre-training of URFM involves representation-based MIM, which distills knowledge from a teacher model to our URFM through MIM. The core principle of MIM approaches, as exemplified by the MAE,[36] is to randomly mask a portion of input image patches and then reconstruct the masked pixels within these patches. In contrast to the original MAE, which focuses on pixel-level reconstruction, our approach leverages representations extracted from the BiomedCLIP[40] as the reconstruction target. BiomedCLIP[40] is a multi-modality biomedical foundation model pre-trained on fifteen million biomedical image-text pairs through contrastive learning, similar to CLIP.[41]

In our approach, the pre-training task for representation-based MIM can be formalized as follows:

$$\min_\theta \mathbb{E}_{\mathcal{I}\sim\mathcal{D}}\mathcal{L}(f_\theta(\mathcal{I}\odot\mathbb{M}),\mathcal{T}(\mathcal{I})\odot(1-\mathbb{M})), \quad \text{(Equation 1)}$$

where $\mathcal{I}$ denotes the input image, $\mathbb{M}$ represents the binary mask indicating which parts of the image should be masked, and $\odot$ signifies the element-wise product operation. Specifically, $\mathcal{I}\odot\mathbb{M}$ represents the image with masked patches, leaving only the unmasked regions visible, while $\mathcal{I}\odot(1-\mathbb{M})$ highlights the masked regions of the image. The function $f_\theta(\cdot)$ refers to the learnable model being pre-trained, $\mathcal{T}$ represents the transformation used to generate the reconstructed target, $\mathcal{L}(\cdot,\cdot)$ is the loss function used for pre-training.

In our implementation, the pre-trained vision encoder from the BiomedCLIP model serves as the transformation $\mathcal{T}$ for generating the reconstructed target. Initially, the entire input image $\mathcal{I}$ is processed through the BiomedCLIP vision encoder to obtain the full-image representation, denoted as $\mathcal{T}(\mathcal{I})$. To focus on reconstructing only the masked portions of the image, we apply the operation

$\mathcal{T}(\mathcal{I}) \odot (1 - \mathbb{M})$ to retain the representation of these masked regions, which acts as the target for reconstruction. Simultaneously, the unmasked patches of the input image, represented as $\mathcal{I} \odot \mathbb{M}$, are fed into the learnable network $f_\theta(\cdot)$. This network adopts an asymmetric encoder-decoder architecture: the encoder is a ViT,[43] and the decoder utilizes a lightweight design consisting of a stack of Transformer blocks[75] followed by a linear projector. This architecture is consistent with the one used in the MAE. The output of this network is the reconstructed representation, denoted as $f_\theta(\mathcal{I} \odot \mathbb{M})$. Finally, the similarity between the target representation $\mathcal{T}(\mathcal{I}) \odot (1 - \mathbb{M})$ and the reconstructed representation $f_\theta(\mathcal{I} \odot \mathbb{M})$ is measured using the loss function $\mathcal{L}(\cdot, \cdot)$. In our case, we employ the mean squared error (MSE) loss to measure the similarity between target and reconstructed representations:

$$\mathcal{L}_{\mathrm{MSE}} = \frac{1}{n} \sum_{i=1}^{n} \left(\mathrm{R}_i^{\mathrm{target}} - \mathrm{R}_i^{\mathrm{recon}} \right)^2 \quad \text{(Equation 2)}$$

where $\mathrm{R}_i^{\mathrm{target}}$ and $\mathrm{R}_i^{\mathrm{recon}}$ represent the target and reconstructed representation of the i-th image, respectively, and n denotes the total number of images used in our pre-training. The primary objective of the pre-training is to minimize the discrepancy between these two representations, thereby aligning them more closely.

The detailed settings of URFM pre-training are summarized in Table 4. Generally, most settings align with those of the MAE, with the exception of the batch size and the total number of pre-training epochs. Specifically, URFM was pre-trained for 400 epochs using a batch size of 1024. This choice of 400 epochs was determined based on the observation that further pre-training did not yield additional improvements in downstream performance. To ensure a fair comparison, these settings are consistent with those used for other baseline models in our study, including MIM-US and Deblur-MIM.

Downstream fine-tuning of URFM

After the pre-training of URFM, only the pre-trained ViT encoder is transferred for downstream fine-tuning. In this study, we adopt full fine-tuning for the downstream transfer, i.e., all parameters of the pre-trained ViT can be adjusted during downstream fine-tuning. The downstream fine-tuning follows a standard supervised learning procedure, where a classification head comprising one single linear layer is appended to the pre-trained ViT encoder to predict classification labels. The cross entropy (CE) loss is employed for the downstream fine-tuning, which is defined as follows:

$$\mathcal{L}_{CE} = -\frac{1}{N} \sum_{n=1}^{N} \sum_{c=1}^{C} y_{n,c} \log(\hat{y}_{n,c}), \quad \text{(Equation 3)}$$

where $y_{n,c}$ represents the ground-truth label for the c-th class of the n-th input image, and $\hat{y}_{n,c}$ denotes the predicted probability for the c-th class of the n-th input image. C is the total number of classes, and N is the total number of images in the corresponding fine-tuning dataset.

The detailed settings of URFM downstream fine-tuning are illustrated in Table 5. Due to the limited number of training images in some downstream fine-tuning experiments (e.g., 10% training data experiment of the liver mass contains only 44 images for training), the batch size for fine-tuning is uniformly set to a relatively small number: 16. For each downstream task, we select the model that performs best on the validation set as the final model for evaluation on the test set. To ensure results robustness, all downstream fine-tuning experiments are repeated 10 times with different random seeds.

Performance evaluation

All experimental performance results in this study are evaluated using three metrics: average F1 score (Average F1), accuracy (ACC), and the area under the ROC curve (AUC). The Average F1 is the arithmetic mean of the F1 scores across all classes, also known as the macro average F1 score. For results that include error bars, these bars represent the 95% confidence intervals (95% CI), with the center of each bar indicating the mean value of the corresponding metric. Statistical significance is assessed using P values calculated from two-sided t-tests.

QUANTIFICATION AND STATISTICAL ANALYSIS

All experimental performance metrics in this study were evaluated using three standard measures: accuracy (ACC), macro-averaged F1 score (Average F1), and the area under the receiver operating characteristic curve (AUC). The Average F1 score represents the arithmetic mean of the F1 scores across all classes.

For results presented with error bars, these bars denote the 95% confidence intervals (95% CI), with the central point indicating the mean value of the corresponding metric. Statistical significance was assessed using two-sided t-tests, with P values reported accordingly.