

A Fully Open and Generalizable Foundation Model for Ultrasound Clinical Applications

Hongyuan Zhang¹, Yuheng Wu^{1,2}, Mingyang Zhao^{3,4,*}, Zhiwei Chen^{1,5}, Rebecca Li⁶, Fei Zhu^{1,*}, Haohan Zhao^{1,2}, Xiaohua Yuan⁷, Meng Yang⁸, Chunli Qiu⁹, Xiang Cong⁹, Haiyan Chen¹⁰, Lina Luan¹¹, Randolph H.L. Wong¹², Huai Liao¹³, Colin A Graham⁶, Shi Chang⁷, Guowei Tao⁹, Dong Yi¹, Zhen Lei^{1,4,14}, Nassir Navab¹⁵, Sébastien Ourselin¹⁶, Jiebo Luo^{1,17}, Hongbin Liu^{1,14,16} and Gaofeng Meng^{1,4,14,*}

¹Center for Artificial Intelligence and Robotics, Hong Kong Institute of Science & Innovation, Chinese Academy of Sciences, Hong Kong, China

²City University of Hong Kong, Hong Kong, China

³State Key Laboratory of Mathematical Sciences, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing, China

⁴University of Chinese Academy of Sciences, Beijing, China

⁵Division of Electronic Engineering, Faculty of Engineering, The Chinese University of Hong Kong, Hong Kong, China

⁶Accident and Emergency Medicine Academic Unit, The Chinese University of Hong Kong, Hong Kong, China

⁷Xiangya Hospital Central South University, Changsha, China

⁸Hunan Frontline Medical Technology Co., Ltd, Changsha, China

⁹Qilu Hospital of Shandong University, Jinan, China

¹⁰Zhongshan Hospital of Fudan University, Shanghai, China

¹¹Shanghai Geriatric Medical Center, Shanghai, China

¹²Division of Cardiothoracic Surgery, Department of Surgery, The Chinese University of Hong Kong, Hong Kong, China

¹³Department of Pulmonary and Critical Care Medicine, The First Affiliated Hospital, Sun Yat-sen University, Guangzhou, China

¹⁴State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Sciences, Beijing, China

¹⁵Computer Aided Medical Procedures, Technical University of Munich, Munich, Germany

¹⁶School of Biomedical Engineering & Imaging Sciences, King's College London, UK

¹⁷Department of Computer Science, University of Rochester, USA

*Corresponding authors

Abstract. The inherent safety and versatility of ultrasound imaging have made it widely accessible in modern clinical settings for disease diagnosis and health management. Artificial intelligence (AI) that can effectively learn ultrasound representations by integrating multi-source data holds significant promise for advancing clinical care. However, the scarcity of large labeled datasets in real-world clinical environments and the limited generalizability of task-specific models have hindered the development of generalizable clinical AI models for ultrasound applications. In this study, we present EchoCare, a novel ultrasound foundation model for generalist clinical use, developed via self-supervised learning on our curated, publicly available, large-scale dataset EchoCareData. EchoCareData comprises 4.5 million ultrasound images, sourced from over 23 countries across 5 continents and acquired via a diverse range of distinct imaging devices, thus encompassing global cohorts that are multi-center, multi-device, and multi-ethnic. Unlike prior studies that adopt off-the-shelf vision foundation model architectures, we introduce a hierarchical classifier into EchoCare to enable joint learning of pixel-level and representation-level features, capturing both global anatomical contexts and local ultrasound characteristics. With minimal training, EchoCare outperforms state-of-the-art comparison models across 10 representative ultrasound benchmarks of varying diagnostic difficulties, spanning disease diagnosis, lesion segmentation, organ detection, landmark prediction, quantitative regression, imaging enhancement and report generation. The code and pretrained model are publicly released, rendering EchoCare accessible for fine-tuning and local adaptation, supporting extensibility to additional applications. EchoCare provides a fully open and generalizable foundation model to boost the development of AI technologies for diverse clinical ultrasound applications.

Contents

1	Introduction	4
2	Results	6
2.1	The largest ultrasound dataset EchoCareData	6
2.2	Architecture and pre-training protocol of the foundation model EchoCare	6
2.3	EchoCare exhibits excellent performance across diverse ultrasound applications	6
2.3.1	Disease diagnostic classification	8
2.3.2	Anatomical segmentation	8
2.3.3	Fetal cardiac organ detection	8
2.3.4	Fetal brain landmark predication	9
2.3.5	Cardiac ejection fraction regression	10
2.3.6	Low-quality imaging enhancement	10
2.3.7	Clinical report generation	10
3	Discussion	12
4	Methods	14
4.1	Model design and pretraining	14
4.1.1	Unified masked pretraining	14
4.1.2	Hierarchical pretraining	14
4.2	Training settings	15
4.3	Data curation for pretraining	15
4.4	Quality control and evaluation of the EchoCareData dataset	15
5	Statistical analysis	18

6 Data collection and analysis 18

7 Data and code availability 19

1 Introduction

Ultrasound imaging stands as a cornerstone of modern medicine, celebrated for its unique combination of real-time assessment, cost-effectiveness, and inherent safety. This non-invasive and radiation-free modality allows for the dynamic visualization of physiological processes, securing its indispensable role in a wide range of clinical practices [28]. Despite these advantages, ultrasound diagnostic is heavily reliant on the skill of the sonographer and the specialized expertise required to interpret the complex, often subtle, visual information. This inherent complexity, coupled with the ubiquity and versatility of ultrasound, has spurred significant interest in leveraging artificial intelligence (AI) to advance its use. As ultrasound imaging expands to new anatomical regions and clinical applications, there is a growing demand for versatile and generalizable AI models that can adapt to diverse clinical tasks and organs with minimal reliance on new labeled data. Meeting this demand will not only broaden the application of ultrasound analysis but also accelerate the deployment of smart healthcare solutions, making high-quality diagnostics more accessible and efficient.

Recent advances in foundation models (FM) using self-supervised learning have opened new frontiers in medical AI [36, 33, 8, 27, 20, 21]. These models learn general-purpose feature representations directly from raw data, eliminating dependence on extensive expert annotations. Upon completion of pretraining, these models can be effectively adapted to a wide array of downstream clinical tasks with minimal or no additional fine-tuning. This paradigm represents a significant advantage over conventional medical AI approaches, which are typically limited to specific anatomical structures or require extensive retraining when adapted to each new clinical application. However, pretraining of foundation models requires large-scale and diverse datasets, making data acquisition and rigorous curation essential for developing clinically reliable and generalizable systems.

Building on the success of vision foundation models, researchers have started adapting these approaches to ultrasound imaging analysis [15, 16, 31]. Although initial results show promise, several critical challenges could limit their potential clinical impact. First, the scale of available ultrasound datasets remains relatively small, undermining the reliability of models for clinical deployment. Moreover, much of the pretraining data employed in previous studies is private, creating barriers to reproducibility, broader research and application. Second, current collections often focus on narrow anatomical regions, which is insufficient to fully capture the diversity of whole-body regions. This limitation restricts their utility in comprehensive clinical workflows. Third, most approaches rely on off-the-shelf vision foundation model frameworks [13], failing to systematically explore network architecture optimizations tailored to the morphological complexity and spatial hierarchies of anatomical structures. This oversight limits the model’s ability to capture anatomical relationships across scales and organs during pre-training. Finally, existing research mainly focuses on a few downstream tasks such as image classification or segmentation, leaving open questions about model capabilities for more diverse clinical applications.

In this work, we introduce EchoCare, a novel foundation model for ultrasound images, accompanied by a systematic investigation of its utility across a diverse spectrum of clinical tasks. EchoCare is pre-trained on EchoCareData, our newly curated large-scale and openly accessible dataset comprising 4.5 million ultrasound images. Collected from multi-center, multi-device, multi-modality, and multi-ethnic global sources, EchoCareData ensures diverse data representation. EchoCareData covers 9 major regions and 52 anatomical organs of the human body, supporting models pretrained on it to generalize effectively across comprehensive whole-body ultrasound clinical applications. We have also optimized the architecture of the vision foundation model to better capture hierarchical anatomical structures, from broad ultrasound regions (*e.g.*, abdomen) to specific organs (*e.g.*, liver, kidney), enabling the model to mimic human-like clinical diagnostic reasoning. Extensive evaluations across eight categories of core ultrasound clinical tasks of varying diagnostic difficulties, such as lesion segmentation, organ detection, disease diagnosis, and quantitative regression, reveal that EchoCare significantly outperforms state-of-the-art general-domain foundation models, underscoring the critical need for ultrasound-specific models. Compared with leading ultrasound-focused foundation models, EchoCare also demonstrated superior performance, highlighting the advantages of pretraining on large, diverse data. We will release both EchoCare and the EchoCareData to promote clinical AI development in ultrasound images upon publication.

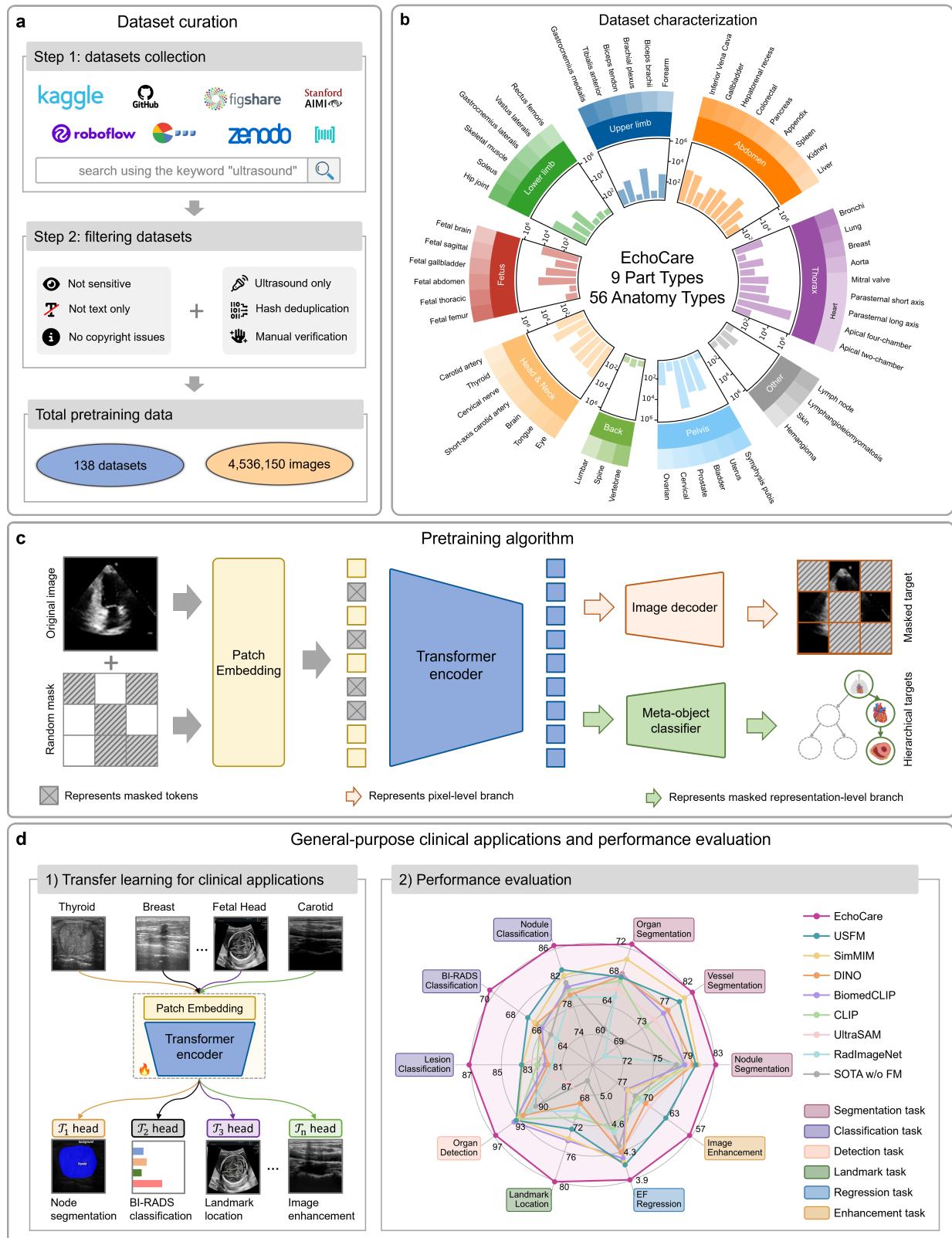


Figure 1 | Overview of this study. Caption on next page.

(Previous page.) **Figure 1: Overview of this study.** **a.** The ultrasound data from over 20 countries and 5 continents are collected, encompassing over 4 million ultrasound images. **b.** The constructed ontology shows a hierarchy of object types that are used to unify semantic concepts across datasets. Bar plots showing the number of images containing that object type. **c.** Flowchart of EchoCare. EchoCare takes a masked image as input and then outputs the reconstructed ultrasound image. To capture both global anatomical contexts and local ultrasound characteristics, EchoCare also incorporates a novel hierarchical classifier branch. **d.** Performance evaluation across diverse clinical applications, spanning different organs and a wide range of diagnostic difficulties. EchoCare achieved state-of-the-art performance across all downstream clinical tasks. Image was created with BioRender.com.

2 Results

2.1 The largest ultrasound dataset EchoCareData

We establish so far the largest public ultrasound image dataset EchoCareData (Fig. 1a,b), integrating 138 ultrasound image datasets from over 20 countries and 5 continents. Encompassing multiple body organs, scanning devices, imaging modalities, and racial backgrounds (Fig. 1b), the dataset is designed to ensure data diversity and enhance the generalization of pretrained models across diverse clinical applications. EchoCareData adheres to rigorous cohort inclusion and exclusion protocols to ensure high quality, including manual removal of sensitive and non-ultrasound images, as well as text cleaning (Fig. 1a). Using a clinical anatomy system, we generated canonical categorical labels for each image. The dataset’s ontology comprises eight representative clinical regions including head, chest, abdomen, limbs, back, fetus, dorsum, pelvis, and an “other” category, with a hierarchical structure spanning 52 meta-object types (*e.g.*, cardiac ventricle) to 56 specific anatomic types (*e.g.*, left cardiac ventricle), mirroring clinical diagnostic workflows. Moreover, an additional manual inspection was performed by randomly sampling 100 images from each class in EchoCareData to validate correctness. In total, EchoCareData comprises over 4.5 million distinct image-class tuples, spanning five imaging modalities (B-mode, CEUS, Dropper, M-mode, and Elastography), establishing it as a large-scale, diverse resource for clinical ultrasound care.

2.2 Architecture and pre-training protocol of the foundation model EchoCare

Building on EchoCareData, we pretrained EchoCare (Fig. 1c), a novel vision foundation model for ultrasound imaging, and applied it to a suite of clinical tasks. EchoCare employs a modular design based on an extended self-supervised Masked AutoEncoder (MAE) architecture for representation learning, comprising an image encoder to encode input ultrasound image features and two decoders: an image decoder to reconstruct images from sparse patches and an anatomy-classifier decoder for joint learning of hierarchical anatomic features (Fig. 1c). Unlike prior medical foundation models that directly adopt off-the-shelf MAE structures or focus solely on local pixel-level prediction, we introduce a novel representation-level prediction branch, the anatomy-classifier, into the MAE framework. This branch learns global and hierarchical anatomical relationships from body regions to organs to anatomic structures, mirroring clinical diagnostic workflows. For example, the anatomy-classifier predicts pathways such as “Thorax→Heart→Apical two-chamber” and “Thorax→Heart→Apical four-chamber”. Leveraging the inherent hierarchical organization of the anatomy system, this high-level classification process evolves naturally without human intervention. By integrating local pixel-level and global representation-level features, EchoCare enhances the encoder’s ability to interpret ultrasound images, thereby boosting downstream clinical applications. In the following sections, we demonstrate its versatility and generalization to diverse ultrasound clinical tasks.

2.3 EchoCare exhibits excellent performance across diverse ultrasound applications

We systematically evaluated EchoCare diagnostic performance on 11 clinical applications across 8 task types (Fig. 2, Fig. 3 and Fig. 4). These datasets cover tasks ranging from binary diagnosis task to multi-class classification, single-class tumor segmentation to abdominal multi-organ segmentation, as well as ultrasound image enhancement, fetal landmark localization, organ detection, cardiac ejection fraction regression, and clinical report generation. We compare EchoCare with previous state-of-the-art (SOTA) task-specific models (w/o FM) and seven representative foundation models: RadImageNet [22], UltraSAM [23], CLIP [25], BiomedCLIP [35], DINO [34], SimMIM [30], USFM [15]. Each model is fully fine-tuned on the

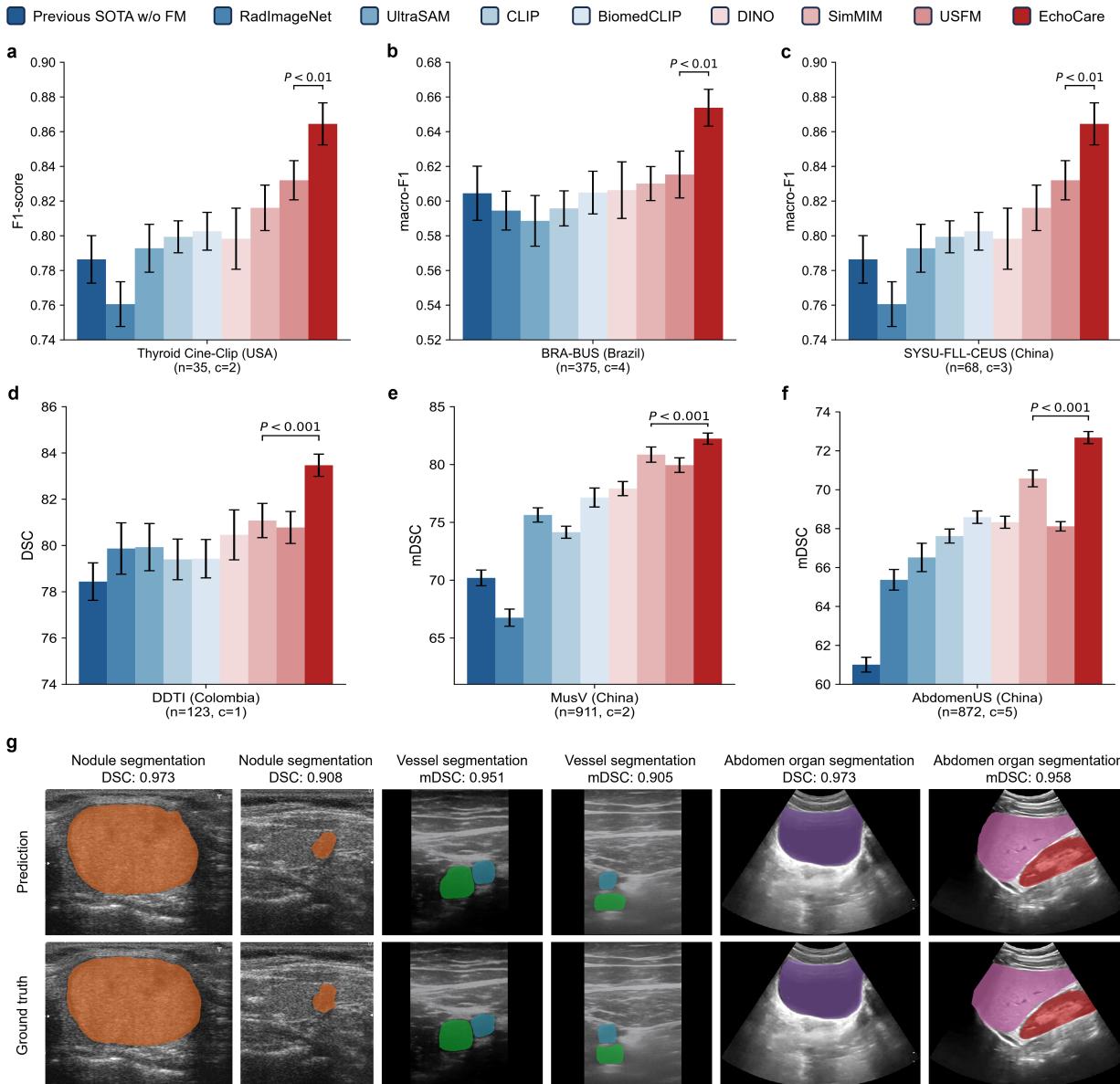


Figure 2 | Evaluation on disease diagnostic classification and anatomical segmentation. **a-f.** EchoCare consistently outperforms previous state-of-the-art (SOTA) models (w/o FM) and other existing foundation models (RadImageNet [22], UltraSAM [23], CLIP [25], BiomedCLIP [35], DINO [34], SimMIM [30], USFM [15]) across different classification and segmentation tasks. Specifically, for classification, we evaluate on benign-malignant classification of thyroid nodules (**a**), breast tumor BI-RADS grading (**b**) and diagnosis of focal liver lesions in abdominal ultrasound (**c**). For segmentation, we evaluate on thyroid node segmentation (**d**), arterial-venous vessel segmentation (**e**), and the abdomen multi-organ segmentation (**f**). The two-sided Wilcoxon signed-rank test was used to assess the statistical differences between EchoCare and the second-best model. **g.** Six examples comparing the segmentation results by EchoCare and the ground truth.

task-specific dataset and evaluated with their corresponding metrics. EchoCare consistently outperformed all other models, achieving significant improvements on 10 clinical tasks. These results validate the effectiveness of Echocare. The domain-specific analyses of the experimental results are as follows.

2.3.1 Disease diagnostic classification

Disease diagnostic classification represents a pivotal clinical application of vision foundation models. High-performance ultrasound foundation models can substantially enhance the accuracy of disease lesion classification, mitigate false-positive decisions, and thereby reduce patient anxiety and costs. To demonstrate the utility in clinical decision-making, EchoCare was validated across three distinct diagnostic classification applications: 1) benign-malignant classification of thyroid nodules; 2) breast tumor BI-RADS grading; and 3) diagnosis of focal liver lesions in abdominal ultrasound.

EchoCare achieved leading performance across all the evaluated classification tasks (Fig. 2a-c). Specifically, EchoCare achieved an AUC (Area Under the ROC Curve) of 86.48% and an F1-score of 87.45% on the thyroid nodule dataset (Fig. 2a); 70.36% accuracy and 65.38% macro-F1 on breast BI-RADS grading (Fig. 2b); and 87.12% accuracy and 83.44% macro-F1 for focal liver lesions (Fig. 2c). Compared with the second-best model (USFM [15]), EchoCare outperformed by average margins of 3.35% (AUC) and 4.25% (F1-score) on thyroid nodules, 3.09% (accuracy) and 3.85% (macro-F1) on breast BI-RADS, and 3.45% (accuracy) and 3.98% (macro-F1) on focal liver lesions. These findings highlight EchoCare as a powerful foundation model capable of learning discriminative image representations, demonstrating great potential in distinguishing subtle differences between hepatocellular carcinoma, hemangiomas, and focal nodular hyperplasia. These lesions often exhibit overlapping sonographic appearances, which is a key challenge in achieving accurate manual diagnosis. Collectively, the experimental results confirm that EchoCare serves as a reliable diagnostic auxiliary tool, advancing ultrasound-based disease diagnostic classification and accelerating the clinical decision-making process.

2.3.2 Anatomical segmentation

Accurate segmentation in ultrasound images enables clinicians to characterize morphological features (*e.g.*, size, shape) and detect pathological abnormalities (*e.g.*, neoplastic lesions), which is fundamental for treatment planning and prognosis assessment. We evaluated different foundation models on three representative ultrasound clinical benchmarks for anatomical segmentation: the DDTI dataset [24] for thyroid node segmentation, the Mus-V dataset [9] for arterial-venous vessel segmentation, and the abdomen multi-organ segmentation.

Compared with existing methods, EchoCare achieved significantly higher performance (Fig. 2d-f), surpassing the next best Dice Similarity Coefficient (DSC) by 2.09% and Normalized Surface Dice (NSD) by 2.26% in the thyroid nodule segmentation task, mDSC by 1.36% and mNSD by 1.03% in the vessel segmentation task (Fig. 2d), mDSC by 2.10% and mNSD by 4.36% in the multi-organ segmentation task (Fig. 2f). In the vascular segmentation task, EchoCare outperforms the second-ranked model (SimMIM) by a remarkable margin (Fig. 2e), which holds clinical significance for real-time vascular interventions (*e.g.*, coronary procedures). Furthermore, EchoCare also surpassed previous SOTA without FM architecture (SwinUNETR [12]) in benchmark evaluations. We showed examples comparing EchoCare segmentation and the ground truth across multiple organs, demonstrating the generalizability of EchoCare (Fig. 2g). The strong performance of EchoCare on segmentation tasks represents a breakthrough for comprehensive abdominal assessments, as clinicians require simultaneous visualization of the liver, pancreas, and kidneys to detect pathological relationships (*e.g.*, liver lesions compressing adjacent organs). Such consistency across single-organ, vascular, and multi-organ tasks underscores, EchoCare's capacity to learn generalizable ultrasound features for effective task adaptation.

2.3.3 Fetal cardiac organ detection

Fetal congenital heart disease (CHD) is a leading cause of infant mortality from birth defects, with an incidence reaching up to 8-10 cases per 1,000 live births [6, 29]. Survival rates, requirement for intensive medical care, and risk of developmental disabilities are contingent on the accuracy and timeliness of diagnosis. Thus, early and precise prenatal sonographic diagnosis of CHD has been shown to reduce the risk of perinatal morbidity and mortality. The four-chamber view in fetal echocardiography is a unique and essential tool for assessing CHD. Diagnosis in this view relies on the cardiothoracic diameter ratio (CTR), a biometric defined as the ratio of thoracic to cardiac short-axis diameters. Therefore, detecting thoracic and cardiac regions from four-chamber echocardiograms is a critical step for CTR analysis and represents a foundational step in CHD

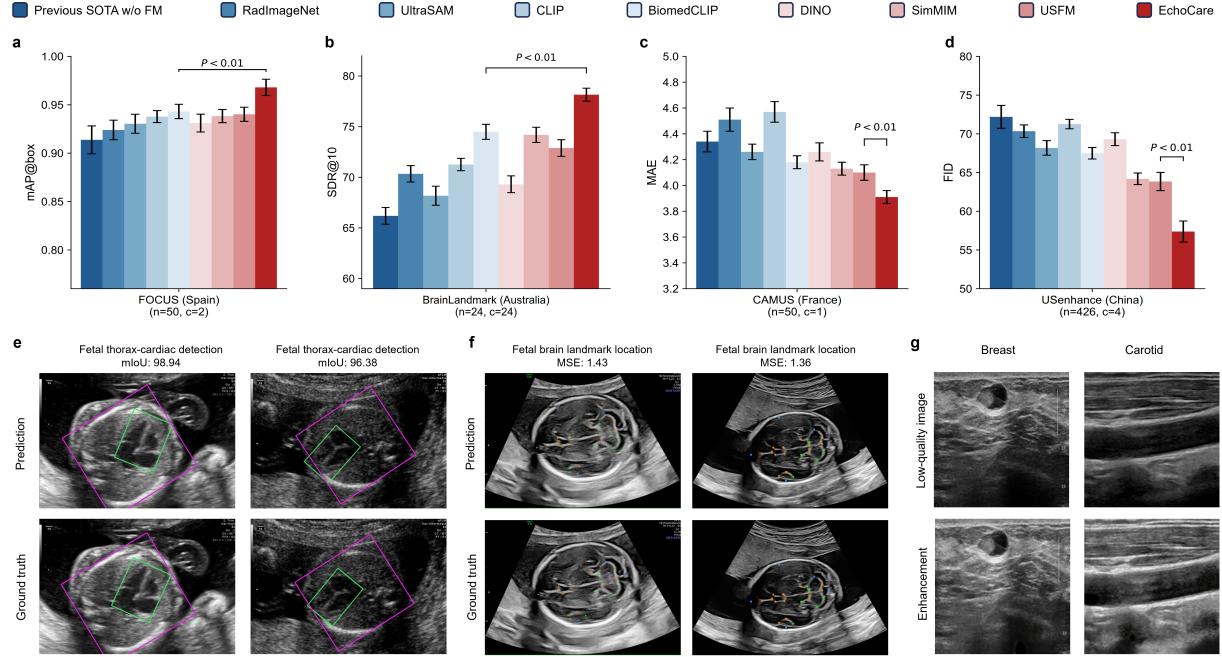


Figure 3 | Evaluation on organ detection, landmark prediction, fraction regression and imaging enhancement. **a-d.** EchoCare consistently outperforms previous SOTA task-specific models (w/o FM) and existing foundation models across different tasks: organ detection (**a**), landmark prediction (**b**), fraction regression (**c**) and imaging enhancement (**d**). The two-sided Wilcoxon signed-rank test was used to assess the statistical differences between EchoCare and the second-best model. **e-f.** Six examples comparing the detection, location and imaging enhancement results by EchoCare and the ground truth.

diagnosis.

In this study, we evaluated the performance of EchoCare against seven state-of-the-art foundation models and one previous SOTA model (Rotated Faster R-CNN [26]) for fetal thorax and cardiac organ detection using the publicly available FOCUS dataset (300 four-chamber fetal echocardiography ultrasound images). EchoCare outperformed all other models significantly (Fig. 3a). Specifically, it 97.26% AP (Average Precision) for thoracic object detection, and 96.11% AP for cardiac object detection. Compared to the top ImageNet-based model (Rotated Faster-RCNN), EchoCare showed even larger margins (5.42% higher mAP). This outcome underscores that ultrasound, specific pretraining, distinct from natural image pretraining, more effectively captures the domain-specific knowledge inherent to ultrasound imaging. Benefiting from its high detection accuracy, EchoCare also ranked first in CTR measurement accuracy (94.42%). We provide examples to visualize the detection results of EchoCare and demonstrate the superior performance by comparing with the ground truth (Fig. 3e). These comprehensive results demonstrate that EchoCare has the potential to enhance and accelerate prognosis prediction for CHD in ultrasound clinical practice.

2.3.4 Fetal brain landmark predication

Brain development involves progressive structural changes from early embryonic stages to several months after birth. Identifying fetal brain structures in ultrasound images enables assessment of cortical and subcortical gray matter changes, serving as a valuable tool for detecting developmental abnormalities. However, manual landmark identification is labor-intensive, time-consuming, and prone to intra- and inter-rater inconsistency.

To address this challenge, we evaluated the performance of EchoCare against other models for predicting fetal brain landmarks using the publicly available BrainBenchmark dataset [7] (104 2D fetal brain ultrasound images acquired at 20-20.6 weeks of gestation from 70 pregnant women). EchoCare outperformed all foundation models significantly (Fig. 3b), achieving a notably lower average MSE (7.71) compared to the second-best model. EchoCare also dominated in successful detection rate (SDR) across all pixel thresholds: at $\tau = 2.0$ pixels, it achieved a SDR of 36.27%, (surpassing the second-best model SimMIM's 30.24); at $\tau = 4.0$ pixels,

it achieved 49.13% (substantially exceeding SimMIM’s 42.87%); and at $\tau = 10.0$ pixels, it achieved 80.16% (versus BiomedCLIP’s 74.49%). We also provide examples to visualize the landmark prediction results of EchoCare and compare with the ground truth (Fig. 3f). These results highlight EchoCare’s superiority in ultrasound-based landmark prediction, positioning it as a promising tool for automated fetal brain assessment.

2.3.5 Cardiac ejection fraction regression

The assessment of Left Ventricular Ejection Fraction (LVEF) is one of the most important manners in the evaluation of cardiac function. It quantifies the proportion of blood ejected from the left ventricle relative to its total end-diastolic volume. In clinical settings, accurate measurement of LVEF is pivotal to the early diagnosis of both congenital and acquired cardiovascular disorders, informs therapeutic decision-making, and enables robust prognostic stratification.

After observing the superior performance of EchoCare across a range of ultrasound clinical tasks, we further evaluated it on the LVEF regression task using the CAMUS benchmark dataset [17]. This dataset encompasses 2D apical four-chamber and two-chamber view sequences from 500 patients. Model performance is quantified by mean absolute error (MAE) with standard error. EchoCare exhibited superior performance and outperformed the other 8 competing approaches (Fig. 3c). It achieved the lowest MAE of 3.91, surpassing the second-best pretrained model (USFM) by a 19% reduction in MAE. Notably, it significantly outperformed the echo-specific state-of-the-art model (EchoMEM), with a significant 43% reduction in MAE. These contributions underscore the potential of EchoCare to advance cardiac LVEF regression and its applicability in real-world clinical workflows.

2.3.6 Low-quality imaging enhancement

High-quality ultrasound imaging is critical for the accurate identification of anatomical structures and disease diagnosis. However, ultrasound examinations using handheld or low-end devices often yield suboptimal images that compromise clinical diagnosis, particularly in resource-limited hospitals or regions. Enhancing such low-quality ultrasound images using AI technologies, for example, through improved contrast, sharpness, and signal-to-noise ratio, alongside noise reduction, could provide a cost-effective alternative to high-end scanners. This approach may also promote the wider adoption of portable ultrasound systems, offering substantial clinical benefits and ultimately improving patient outcomes.

We evaluated EchoCare on the low-quality ultrasound image enhancement task using the USenhance benchmark dataset [10], which encompasses real-world clinical scans from 109 patients across five anatomical regions: thyroid, kidney, liver, breast, and carotid artery. EchoCare was compared with 8 models, including previous SOTA model (EnlightenGAN [14]), ultrasound-based models (RadImageNet [22], UltraSAM [23]), image-text multimodal models (CLIP [25], BiomedCLIP [35]), and self-supervised frameworks (DINO [34], SimMIM [30], USFM [15]). Consistent with previous findings, EchoCare outperformed all competing models across four metrics: NIQE, BRISQUE, PIQE, and FID (Fig. 3d). Specifically, EchoCare achieved mean NIQE, BRISQUE, PIQE, and FID values of 6.35%, 17.62%, 30.16%, and 57.38%, respectively. These visualizations (Fig. 3g) further demonstrate the superior image quality enhancement ability of EchoCare. These results demonstrate that EchoCare can effectively enhance low-quality ultrasound images, highlighting the potential of AI for practical clinical applications in resource-limited settings.

2.3.7 Clinical report generation

Report generation is essential for healthcare system, providing critical information to clinicians and patients for the diagnosis, prognosis, and treatment planning of a wide range of medical applications. Traditionally, ultrasound reports are written manually by sonographers, which is time-consuming and prone to inter-observer variability. Recent advancements in natural language processing and medical image analysis have enabled the development of automated ultrasound report generation systems.

To evaluate the effectiveness of our developed foundation model in ultrasound report generation, we integrate EchoCare into an existing Transformer-based encoder-decoder report generator, where the input is the global visual features extracted from ultrasound images. The integrated model is then fine-tuned on the USData Liver dataset [18], which contains paired ultrasound images and corresponding expert-written reports. The

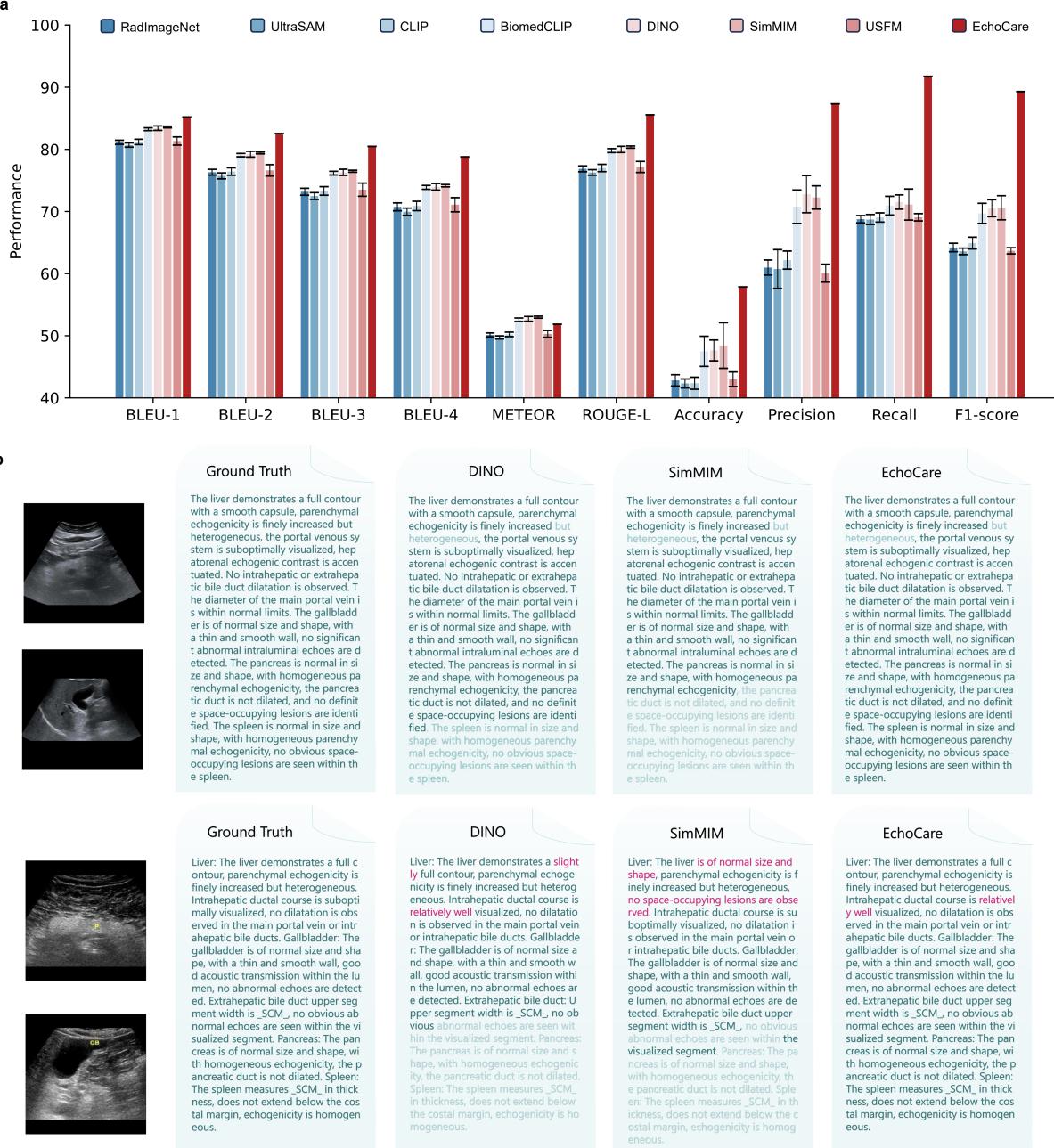
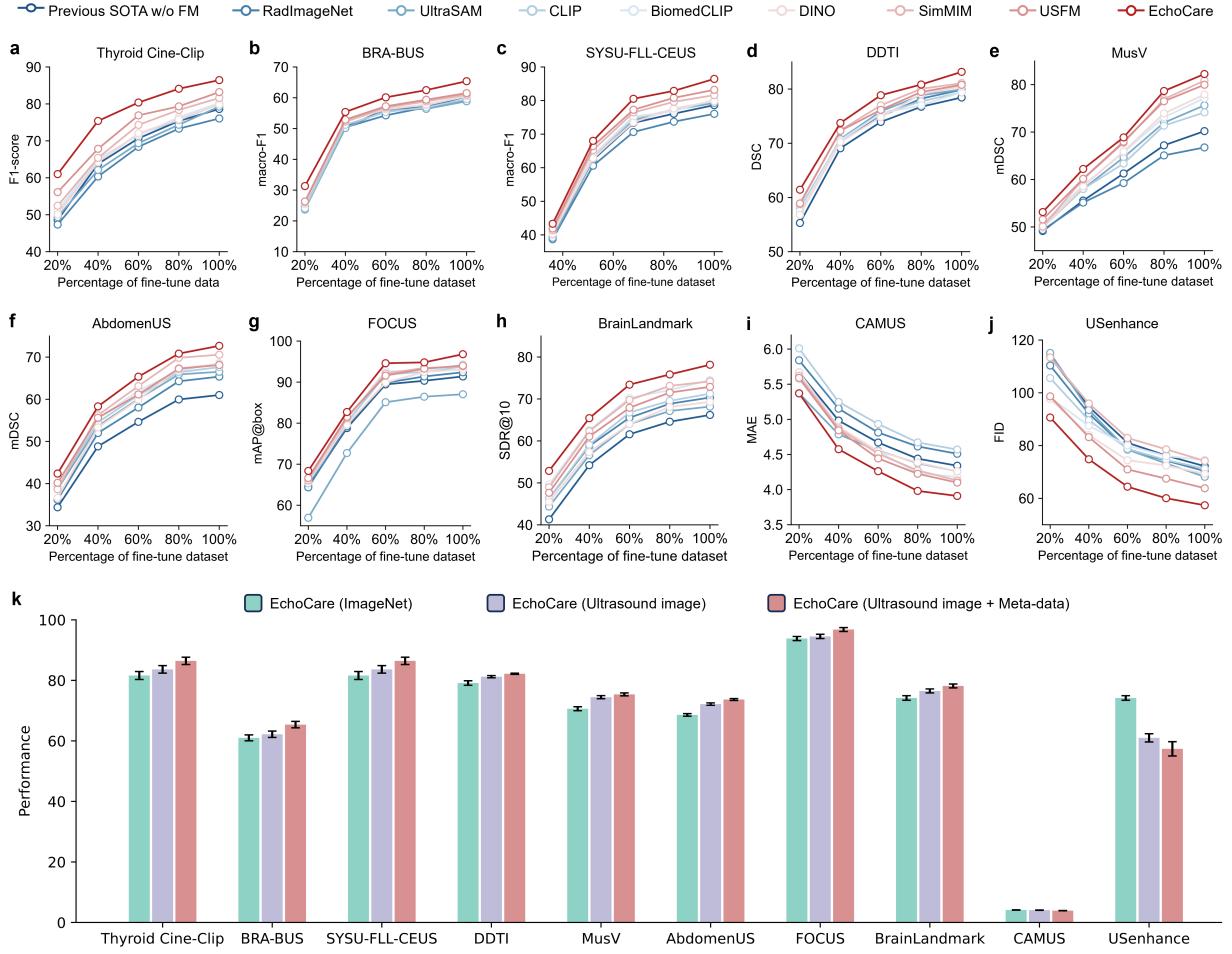


Figure 4 | Clinical report generation on USData [18] Liver dataset. a. Performance (%) comparison of models on the USData [18] Liver dataset using six language metrics (BLEU-1 to BLEU-4, METEOR, and ROUGE-L) and four classification metrics (Accuracy, Precision, Recall, and F1-score). Error bars denote standard deviation across multiple runs. **b.** Example reports generated by the two strongest baseline models (DINO [34] and SimMIM [30]) and EchoCare, compared against ground truth reports. Deep blue text indicates exact matches, light-colored text denotes missing segments, and vivid purple highlights over-generated content.

experimental results (Fig. 4a) demonstrate that EchoCare achieved the best performance across all ten metrics. Notably, compared with the second-best model (SimMIM [30]), EchoCare outperformed by large margins of 4.64% (BLEU-4), 9.43% (accuracy) and 18.70% (F1-score). Besides, the examples of generated reports (Fig. 4b) valid the ability of EchoCare for ultrasound report generation, demonstrating the potential of EchoCare in improving the efficiency, consistency, and accessibility of automatic clinical report generation.



3 Discussion

Ultrasound imaging is a crucial tool in modern medicine. This work presents a novel, open-source foundation model named EchoCare to advance general-purpose clinical ultrasound applications. The model is pre-trained on our curated, publicly available dataset of 4.5 million ultrasound images, featuring a highly diverse and whole-body images sourced from over 20 countries and 5 continents. To evaluate EchoCare's clinical utility, we conducted comprehensive validations across a wide range of downstream ultrasound tasks (lesion segmentation, disease diagnosis, organ detection, landmark prediction, quantitative regression, imaging enhancement and report generation). The results demonstrate the strong effectiveness and generalization capabilities of EchoCare: it consistently outperforms state-of-the-art foundation models such as UltraSAM [23], BiomedCLIP [35], and USFM [15] across all tasks.

The pre-training of our foundation model is powered by EchoCareData, the largest publicly available ultrasound image dataset to date, featuring over 4.5 million images. The vast scale and diverse of this dataset are crucial to our success. The core idea behind our data strategy is simple yet powerful: by aggregating numerous public datasets, we can significantly increase data size, expand protocol coverage, and diversify patient populations. This approach allows our model to learn from a broad spectrum of global sources. This extensive pre-training

also grants EchoCare a remarkable degree of label efficiency for various downstream clinical tasks (Fig. 5a-j), thereby alleviating the substantial annotation workload for medical experts. For instance, in thyroid nodule segmentation, EchoCare can outperform other models using only 80% of the labeled training data. Furthermore, EchoCare showed consistently high adaptation efficiency, suggesting that EchoCare required less time in adapting to downstream clinical applications, *e.g.*, EchoCare can potentially save about 20% ~ 40% of the training time required to achieve convergence for the task of disease prediction.

In addition to the substantial size and broad diversity of EchoCareData, the designed dual-branch architecture further contributes to the superior performance of EchoCare across a wide range of downstream clinical tasks. Unlike previous medical foundation models that relied on standard MAE structures, our enhanced MAE architecture incorporates a unique anatomy-classifier branch. This branch is designed to learn global and hierarchical anatomical relationships, mirroring a clinician’s diagnostic process. By integrating this high-level, representation-based learning with the local, pixel-level prediction of the MAE, our model’s encoder gains a deeper understanding of ultrasound images. This dual-learning approach significantly boosts the model’s ability to interpret images and perform well in a wide range of downstream clinical applications (Fig. 5k).

While EchoCare has demonstrated promising potential of pretrained foundations for ultrasound analysis, several methodological frontiers remain. First, current pretraining exclusively uses image data, omitting clinically actionable text modalities (*e.g.*, ultrasound diagnostic reports). Future iterations will integrate vision-language learning through curated datasets, enabling joint modeling of ultrasound images and associated clinical narratives to expand clinical applications. Second, EchoCare currently treats dynamic modalities (*e.g.*, videos) to static frames, thus failing to utilize the temporal cues essential for applications like cardiac motion analysis or vascular flow assessment. We will extend the architecture to incorporate spatio-temporal transformers, enabling end-to-end training on native video sequences and preserving temporal dynamics. Third, although results across 10 downstream clinical tasks demonstrate translational potential, rigorous validation is required before clinical adoption, such as real-world integration with clinical decision support systems.

In conclusion, we introduce EchoCare, a novel vision foundation model for ultrasound analysis, pretrained on our curated EchoCareData, which comprises over 4.5 million ultrasound images and is the largest ultrasound dataset to date. By integrating a novel architecture with a massive, diverse dataset, EchoCare establishes an efficient new paradigm for ultrasound image analysis, demonstrating robust adaptability to a broad spectrum of clinical ultrasound tasks and delivering significant performance gains over existing foundation models. Critically, we have made both the EchoCare model and EchoCareData publicly accessible to accelerate advancements in medical AI, improving clinical decision-making and patient care.

4 Methods

4.1 Model design and pretraining

For large-scale visual pretraining on EchoCareData, we proposed EchoCare, a self-supervised framework for pre-training large vision transformer architectures based on the Masked Image Modeling (MIM) paradigm. Specifically, EchoCare adopts a modular design, comprising an image encoder, an image decoder, and a meta-object classifier (Fig. 1c), each module described in detail below.

The input to EchoCare is a masked image, which is passed along to the image. The image encoder processes the high-resolution image and outputs multi-scale downsampled embeddings. We provide a flexible choice of backbone architectures with Swin Transformer base and large versions. The image decoder outputs a reconstructed image that has the same size as the original image, with a grayscale value between 0 and 1 for each pixel. The meta-object classifier includes input from the image and output object semantics. The output object semantics includes three levels: part, organ and anatomical structure. We follow SimMIM and SwinUNETR to build the image decoder head. The decoder is a transformer that gradually upsamples the image features back to high-resolution pixels. At the last layer, the attention dot product on the pixel embeddings delivers the reconstructed image.

4.1.1 Unified masked pretraining

The input image $x \in \mathbb{R}^{H \times W \times C}$ was split into N image patches $\{x_i^p\}_{i=1}^N$ and then tokenized into $z = [z_1, \dots, z_N] \in \mathbb{V}^{h \times w}$ as the output labels of MIM using an image patch embedding layer. At the input layer, 50% image patches were randomly masked, and then the model predicted the visual tokens z_i of the masked patches. Next, we replaced the masked patches with a learnable embedding $e_{[M]} \in \mathbb{R}^D$, making the input corrupted image patches $x^M = \{x_i^p : i \notin M\}_{i=1}^N \cup \{e_{[M]} : i \in M\}_{i=1}^N$ that are fed into the transformer encoder. To optimize the model, we employ a reconstruction loss that aims to minimize the difference between the predicted pixel values of the masked image patches, \hat{x}_i^p , and the ground truth pixel values, x_i^p . Specifically, the reconstruction loss is defined as the Mean Absolute Error (MAE) between the predicted and original patches:

$$\mathcal{L}_{\text{MIM}} = \frac{1}{M} \sum_{i \in M} |\hat{x}_i^p - x_i^p|. \quad (1)$$

4.1.2 Hierarchical pretraining

The second pre-trained output (*i.e.*, the meta-object classifier) is used to further train EchoCare to represent images using hierarchical learning. Therefore, we designed a hierarchical loss for image global representation learning. Specifically, let's assume there are N_p body parts at the first level, which encompass N_o organs at the second level. Based on these N_o organs, there are N_a anatomical structures at the third level. Hence, the meta-object classifier has $N_p + N_o + N_a$ outputs. For each category, if a class is labeled positive, all its ancestor nodes (*i.e.*, superclasses) should be labeled positive. And, if a class is labeled negative, all its child nodes (*i.e.*, subclasses) should be labeled negative. To ensure the satisfaction of the above hierarchy constraints, we estimate a hierarchy-coherent score vector $P \in [0, 1]^{N_p + N_o + N_a}$. For class i , the updated score vector $p = [p_i] \in [0, 1]$ in P is given as:

$$\begin{cases} p_H = \min(s_u) & \text{if } \hat{l} = 1, \\ 1 - p_H = \min(1 - s_u) = 1 - \max(s_u) & \text{if } \hat{l} = 0. \end{cases} \quad (2)$$

Thus, after getting the hierarchical probabilities, we could maximize the log-likelihood between the probabilities and ground truth classification labels:

$$\mathcal{L}_{\text{HIE}} = \sum -\hat{l} \log(p_H) - (1 - \hat{l}) \log(1 - p_H). \quad (3)$$

4.2 Training settings

Image augmentations included random vertical flip ($P = 0.5$), random horizontal flip ($P = 0.5$), and random crop ($P = 0.5$) to convert images to greyscale and weak colour jittering ($P = 0.2$) with specific adjustments to brightness, contrast, saturation and hue. We pretrained EchoCare for one million steps using the pretraining loss of $\mathcal{L}_{\text{MIM}} + \mathcal{L}_{\text{HIE}}$ for images. The batch sizes were 256, and EchoCare used an input image with 256×256 pixels and then patched as 2×2 pixels. We used the AdamW optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.9$ and $\epsilon = 0.9$ for optimization. We used a cosine learning rate decay scheduler with a peak learning rate of 1.0×10^{-4} and a linear warmup of 10,000 steps. The weight decay was set as 0.05, and the stochastic depth with a rate of 0.1 was used.

4.3 Data curation for pretraining

Our data curation process commenced with a systematic search of open academic repositories, including Zenodo [5], Mendeley [4], Stanford AIMI, Figshare, and data/code platforms such as Kaggle [3], GitHub [1], and medical challenge portals (*e.g.*, Grand Challenge [2]). All data collection was concluded by 1 March 2025. Using "ultrasound" as a keyword, we retrieved approximately 13,000 potential datasets for initial screening. The raw dataset underwent a series of exclusion steps: 1) datasets were filtered to retain common file formats—including image files (*e.g.*, PNG, JPG, BMP) and compressed archives (*e.g.*, ZIP, RAR, TFRecord)—to confirm the presence of ultrasound images; 2) GPT-4o was utilized to extract direct download links from dataset descriptions in excluded text-only candidates; 3) preliminary deduplication was performed by comparing download URLs and computing image hash values for efficiency; and 4) manual curation was implemented to eliminate intra-organ redundancy through fine-grained filtering. Moreover, to mitigate intrinsic anatomical sampling biases and ensure comprehensive coverage, we strategically prioritized underrepresented anatomical structures through targeted efforts: submitting formal access requests to specialized repositories (*e.g.*, EchoNet-Dynamic) and directly contacting authors of ultrasound studies to procure supplementary datasets. Following our rigorous inclusion-exclusion pipeline, we compiled 138 high-quality ultrasound datasets comprising over 4.5 million images, spanning nine major anatomical regions and 32 representative organs.

4.4 Quality control and evaluation of the EchoCareData dataset

An additional quality control and evaluation pipeline was implemented during construction of the EchoCareData dataset. To ensure data integrity, ultrasound images underwent a rigorous purification workflow: (1) Removal of extraneous patient metadata surrounding the image; (2) Discarding of completely empty images or those containing fewer than 1,000 valid (non-zero) pixels; (3) For ultrasound videos, systematic uniform sampling at 10-frame intervals to mitigate redundancy. Post-hoc evaluation of the filtering and deduplication processes was conducted as follows: after data filtering, a random sample of 100 excluded candidates was analyzed, confirming no valid ultrasound images or additional data links. Following deduplication, 100 potential duplicate datasets were manually assessed using a predefined similarity threshold ($\geq 95\%$), verifying their redundancy. These procedures streamlined the dataset from 1,136 to 334 entries by eliminating redundancy. With the inclusion of specialized anatomically balanced datasets, we ultimately curated 138 high-quality ultrasound datasets.

External validation tasks and benchmark datasets

To validate the generalizability of the pretrained foundation model EchoCare, we established 10 external validation tasks spanning representative clinical ultrasound scenarios including lesion segmentation, disease diagnosis, one-shot recognition, and quantitative regression. These tasks leveraged independent datasets covering anatomical regions such as thyroid, venous systems, abdominal organs, and cardiac structures. All external datasets were explicitly excluded from the EchoCareData pretraining corpus to prevent data leakage and ensure unbiased evaluation of pretraining effects. Below is a detailed breakdown of each clinical validation task and corresponding dataset, organized by task category to highlight translational relevance.

Thyroid node segmentation on DDTI dataset (1 classes): The DDTI dataset [24] for thyroid node segmentation comprises 388 patients with B-mode ultrasound scans from the Instituto de Diagnóstico Médico S.A. and National University of Colombia, annotated for nodule lesion segmentation. Images were extracted

Table 1 | Summary of the 10 clinical applications and the dataset distributions.

Name	Challenge	Task type	Country	Anatomy part	Anatomy organ	Train	Test
DDTI [24]	Thyroid node segmentation	Node segmentation	Colombia	Head & Neck	Thyroid nodule	522	123
MusV [9]	Vessel segmentation	Vessel segmentation	China	Head & Neck, Low limb	Carotid and femoral vessels	2,203	911
AbdomenUS	Abdomen organ segmentation	Organ segmentation	China	Abdomen	Liver, Kidney, Pancreas, Bladder, Spleen	3,345	872
Thyroid Cine-Clip [32]	Thyroid node diagnose	Node classification	USA	Head&Neck	Thyroid nodule	157	35
BRA-BUS [11]	BI-RADS category assessment	BI-RADS classfication	Brazil	Thorax	Breast	1,500	375
SYSU-FLL-CEUS [19]	Liver lesion recognition	Lesion classification	China	Abdomen	Liver	285	68
FOCUS [29]	Thorax and cardiac organ detection	Organ detection	Spain	Fetus	Fetal thorax and cardiac	250	50
BrainLandmark [7]	Brain landmark location	Landmark location	Australia	Fetus	Fetal brain	80	24
CAMUS [17]	Ejection fraction regression	EF regression	France	Thorax	Heart	450	50
USenhance [10]	Low-quality image enhancement	Image enhancement	China	Head & Neck, Abdomen, Thorax	Thyroid, Kidney, Liver, Breast, Carotid artery	1,654	426
USreport [18]	Clinical report generation	Report generation	China	Head & Neck, Abdomen, Thorax	Thyroid, Liver, Breast	1,118	279

from thyroid ultrasound video sequences acquired using TOSHIBA Nemio 30 and Nemio MX systems, equipped with 12 MHz convex and linear transducers. Accurate automated segmentation of thyroid nodules enables clinicians to assess morphological features—including size, shape, and margins—to discriminate between benign and malignant lesions, which is critical for early thyroid disease diagnosis. Sub-images were cropped from 42 composite sequences and integrated with single-frame images, yielding a total of 645 ultrasound images with an average resolution of 348×280 pixels and mean mask area of 153.25 pixels. For cross-validation, data were split at the patient level in an 8:2 ratio, resulting in 308:80 patient folds (522:123 images) for training and evaluation.

Artery&vein segmentation on Mus-V dataset (2 classes): The Mus-V dataset [9] for vascular segmentation comprises 3,114 ultrasound images from the Institute of Automation, Chinese Academy of Sciences, annotated for carotid and femoral vessel segmentation. Images were acquired from 11 healthy volunteers using an Angel Pionner H20 Ultrasound Scanner, capturing carotid and femoral vessels in the arm and neck regions. Accurate arterial-venous segmentation is critical for real-time low-risk vascular interventions—such as those for coronary and peripheral vascular diseases—enabling clinicians to precisely target vessels and minimize the risk of adjacent structure injury. The dataset includes separate annotations for arteries and veins to facilitate vascular analysis and identification, with images sampled from 105 videos (5-160 frames per video) at 400×600 pixel resolution. For evaluation, official train-test splits were used to achieve an 8:2 patient-level division, yielding 2,203:911 images for training and validation.

Abdominal multi-organ segmentation on AbdomenUS dataset (5 classes): Beyond single/two-class segmentation, EchoCare was further validated on multi-organ segmentation to demonstrate its potential in reducing annotation burdens on experts. The AbdomenUS dataset for multi-organ segmentation encompasses 4,217 ultrasound images from BGI Genomics Co., Ltd., acquired from 64 volunteers using the MGUS-R3 ultrasound system. Images were annotated for at least one of five abdominal organs: 1) liver, 2) pancreas, 3) kidney, 4) bladder, and 5) spleen. This multi-organ annotation framework allows clinicians to systematically evaluate anatomical morphology—including organ shape, positional relationships, and pathological signs—from diverse sonographic perspectives. For model training and validation, data were divided into an 8:2 ratio at the case level, yielding a training set of 51 cases (3,345 B-mode images) and a validation set of 13 cases (872 B-mode images).

Thyroid nodule false positive mitigation on ultrasound cine-clip dataset (2 classes): The thyroid nodule false positive mitigation task leverages the Ultrasound Cine-clip dataset [32] from the Center for Artificial Intelligence in Medicine & Imaging, comprising 192 histopathologically confirmed thyroid nodules

(175 benign, 17 malignant) across 167 patients (mean age 56 ± 16 years, 137 female) who underwent cine ultrasound between April 2017 and May 2018. The dataset includes ultrasound cine-clip sequences, radiologist-annotated segmentation, patient demographics, lesion metrics (size/location), and definitive histopathological diagnoses. Given the nonspecific nature of ultrasound findings, which often lead to unnecessary biopsies, AI-driven pre-biopsy triage of benign and malignant nodules holds significant clinical value for reducing false positive cancer classifications. All ultrasound acquisitions were performed using Logiq E9 (GE Healthcare) or Siemens S2000 systems, with images obtained by certified sonographers from supine patients with slightly hyperextended necks. The cine-clips feature 802×1054 pixel resolution. Following official dataset splits, the cohort was partitioned into training (157 cine-clips, 4/5) and validation (35 cine-clips, 1/5) subsets to ensure reproducible evaluation.

BI-RADS category assessment on BRA-BUS dataset (4 classes): The Breast Imaging Reporting & Data System (BI-RADS) category assessment leverages the BRA-BUS dataset [11], which offers a standardized lexicon and reporting framework for breast ultrasound. BI-RADS facilitates consistent communication of imaging findings among radiologists and clinicians, with final assessments categorized by malignancy likelihood: categories 2 (benign), 3 (probably benign), 4 (suspicious), and 5 (highly suggestive of malignancy), as annotated by senior ultrasonographers. The BRA-BUS dataset comprises 1,875 anonymized images from 1,064 female patients, acquired using four ultrasound systems (GE Logiq 5, GE Logiq 7, Toshiba Aplio 300, GE U-Systems) with linear-array transducers at the National Institute of Cancer (Rio de Janeiro, Brazil). For validation, an official 5-fold cross-validation strategy was employed, combining four folds into the training set (800 patients, 1,500 images) and using the remaining fold for validation (264 patients, 375 images).

Focal liver lesion diagnosis on SYSU-FLL-CEUS dataset (3 classes): The focal liver lesion (FLL) diagnosis task leverages the SYSU-FLL-CEUS dataset [19], encompassing contrast-enhanced ultrasound data for three pathological types: 186 hepatocellular carcinoma (HCC), 109 hemangioma (HEM), and 58 focal nodular hyperplasia (FNH) cases. Acquired from the First Affiliated Hospital of Sun Yat-sen University using an Aplio SSA-770A ultrasound system (Toshiba Medical Systems), the dataset captures FLLs with heterogeneous patterns, varying in size, contrast intensity, morphological features, and anatomical location (resolution: 768×576 pixels). Early FLL characterization from ultrasound is critical for timely oncological intervention, as these lesions exhibit diverse imaging phenotypes. The dataset was case- and label-stratified into 8:2 training-evaluation folds to maintain class distribution: the training set includes 150 HCC, 88 HEM, and 47 FNH cases, while the evaluation set contains 36 HCC, 21 HEM, and 11 FNH cases.

Fetal thorax and cardiac detection on FOCUS dataset (2 objects): The FOCUS dataset [29] is designed for fetal thorax and cardiac organ detection, comprising 300 four-chamber view fetal echocardiography ultrasound images from 217 subjects across Hospital Clinic and Hospital Sant Joan de Deu in Barcelona, Spain. This dataset captures the cardiothoracic diameter ratio—a critical biometric for assessing fetal congenital heart disease—via ellipse annotations of cardiac and thoracic regions in every image. All images (230×245 pixels, uniform resolution) feature distinct annotations for fetal cardiac and thoracic structures, varying in size, aspect ratio, and rotational orientation. Following official patient-level splits to prevent data leakage, the dataset was partitioned into 250 training images and 50 evaluation images, maintaining clinical representativeness.

Brain landmark detection on BrainBenchmark dataset (24 landmarks): The brain landmark detection task leverages the BrainBenchmark dataset [7], comprising 104 2D fetal brain ultrasound images acquired at 20–20.6 weeks of gestation. Developed for monitoring neurodevelopmental trajectories, this benchmark captures structural changes from embryonic stages to postnatal development, with images obtained from 70 pregnant women (median age 31 years, range 18–42) via routine mid-trimester scans using a Voluson E10 ultrasound system with a high-frequency transabdominal probe (C2-9). Each image is annotated with 24 anatomical landmarks including 4 skull landmarks, 3 thalamic landmarks, 8 cerebellar perimeter landmarks, 4 cavum landmarks, 3 Sylvian fissure landmarks, and 2 midline edge landmarks. Images were collected from 70 subjects with variable scanning frequencies (8 women scanned three times, 18 women twice, and 44 women once), all without detected abnormalities. For validation, an 8:2 image-level split yielded 80 training and 24 evaluation images to ensure developmental stage representativeness.

Ejection fraction prediction on CAMUS dataset (500 cases): The CAMUS dataset [17] for ejection fraction prediction comprises 500 2D ultrasound sequences, recognized as a standard benchmark for cardiac function assessment. This regression task involves inputting ultrasound frame sequences to predict left ventricular ejection fraction (LVEF), a critical biomarker for evaluating cardiac health and diagnosing heart disease, particularly when derived from four-chamber view acquisitions. Ultrasound sequences were acquired using GE Vivid E95 scanners (GE Vingmed Ultrasound, Horten, Norway) with a GE M5S probe (GE Healthcare, US) at the University Hospital of St Etienne (France). Each sequence includes manual annotations of left ventricular volumes at end-diastole and end-systole, from which ejection fraction is calculated. Following official protocols, the dataset was partitioned into 450 training and 50 validation cases to ensure reproducible evaluation of LVEF prediction models.

Image enhancement based on USenhance dataset (5 organs): The ultrasound image enhancement task [10] leverages the USenhance Challenge 2023 dataset, comprising 2,100 ultrasound images (1,050 unpaired low/high-quality image pairs) across five organs (thyroid, kidney, liver, breast, and carotid artery) from 109 patients. AI-driven enhancement of high-quality ultrasound images from low-fidelity inputs obviates the need for hardware upgrades, driving technological innovation in ultrasound devices and enabling more precise clinical applications. The dataset includes images acquired using diverse imaging systems: thyroid imaging employs the mSonics MU1 (low-end) and Toshiba Aplio 500 (high-end); carotid artery and abdominal imaging use SSUN (low-end) and Toshiba Aplio 500 (high-end); breast imaging utilizes the mSonics MU1 (low-end) and Aixplorer system from SuperSonic Imaging (high-end). All images were resized to a uniform 256×256 pixel resolution. Following an organ-stratified 8:2 split, the dataset was partitioned into 837 training image pairs (232 thyroid, 161 breast, 97 kidney, 119 liver, 228 carotid) and 213 validation image pairs (59 thyroid, 41 breast, 25 kidney, 30 liver, 58 carotid), ensuring clinical representativeness across anatomical structures.

Ultrasound text report generation on USreport dataset (3 organs): The USreport dataset [18] is designed for ultrasound text report generation, comprising three independent clinical corpora of ultrasound image-text pairs covering breast, thyroid, and liver examinations. Specifically, it includes 3,534 breast, 2,460 thyroid, and 1,397 liver cases, all sourced from the ultrasonic department database of the PLA General Hospital. AI-driven automated report generation from ultrasound images holds promise to streamline clinical diagnostic workflows. Each report is associated with two representative images selected by clinicians, forming image-text pairs for model training. Following official data splits, an 8:2 train-validation partition was applied: Liver: 1,118 training, 279 validation cases; Breast: 2,827 training, 707 validation cases; Thyroid: 1,968 training, 492 validation cases.

5 Statistical analysis

For all experimental results, performance metrics are reported as mean \pm standard deviation across 20 independent trials. For each evaluation task, two-sample t-tests were conducted between the best-performing model and all others, with statistical significance denoted by asterisks ($*p < 0.05$, $**p < 0.01$, $***p < 0.001$). A two-sided P-value < 0.05 was considered statistically significant. For the thyroid nodule false positive mitigation binary-classification task, accuracy, sensitivity, and specificity were determined using the optimal cut-off value derived from the ROC curve to maximize the Youden index (sensitivity + specificity - 1). All statistical analyses were performed using Python (version 3.10) and MedCalc (version 22.032). Across all experimental settings, results are visualized via box plots (version 3.9.1) showing quartiles and whiskers at $1.5 \times$ interquartile range, based on 20 repeated runs to characterize model performance variability.

6 Data collection and analysis

All source image data were from publicly available datasets. We used Python (version 3.10) to curate and preprocess the image.

7 Data and code availability

The curated ultrasound images in EchoCareData are available at <https://echocare.cares-copilot.com/>. All public datasets used in this work are listed with detailed download links on the project homepage. The pretrained foundation model EchoCare is publicly available at <https://github.com/CAIR-HKISI/EchoCare>, including source code, installation guidelines, model weights, example datasets, and downstream task evaluation scripts.

References

- [1] Github repository. <https://github.com/>.
- [2] Grand-challenge platform. <https://grand-challenge.org/>.
- [3] Kaggle. <https://www.kaggle.com/datasets/bachaboos/tf-for-pocovid-ultrasound>.
- [4] Mendeley. <https://www.mendeley.com/>.
- [5] Zenodo. <https://zenodo.org/>.
- [6] Nathalie Jeanne Bravo-Valenzuela, Alberto Borges Peixoto, and Edward Araujo Júnior. Prenatal diagnosis of congenital heart disease: A review of current knowledge. *Indian heart journal*, 70(1):150–164, 2018.
- [7] Mariano Cabezas, Yago Diez, Clara Martinez-Diago, and Anna Maroto. A benchmark for 2d foetal brain ultrasound analysis. *Scientific Data*, 11(1):923, 2024.
- [8] Richard J Chen, Tong Ding, Ming Y Lu, Drew FK Williamson, Guillaume Jaume, Andrew H Song, Bowen Chen, Andrew Zhang, Daniel Shao, Muhammad Shaban, et al. Towards a general-purpose foundation model for computational pathology. *Nature medicine*, 30(3):850–862, 2024.
- [9] Yimeng Geng, Gaofeng Meng, Mingcong Chen, Guanglin Cao, Mingyang Zhao, Jianbo Zhao, and Hongbin Liu. Force sensing guided artery-vein segmentation via sequential ultrasound images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 656–666. Springer, 2024.
- [10] Yi Guo, Shichong Zhou, Jun Shi, and Yuanyuan Wang. Ultrasound image enhancement challenge 2023. In *International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI) 2023*. Zenodo, 2023.
- [11] Wilfrido Gómez-Flores, Maria Gregorio-Calas, and Wagner Pereira. Bus-bra: A breast ultrasound dataset for assessing computer-aided diagnosis systems. *Medical Physics*, 51, 11 2023.
- [12] Ali Hatamizadeh, Vishwesh Nath, Yucheng Tang, Dong Yang, Holger R Roth, and Daguang Xu. Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In *International MICCAI brainlesion workshop*, pages 272–284. Springer, 2021.
- [13] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022.
- [14] Yifan Jiang, Xinyu Gong, Ding Liu, Yu Cheng, Chen Fang, Xiaohui Shen, Jianchao Yang, Pan Zhou, and Zhangyang Wang. Enlightengan: Deep light enhancement without paired supervision. *IEEE transactions on image processing*, 30:2340–2349, 2021.
- [15] Jing Jiao, Jin Zhou, Xiaokang Li, Menghua Xia, Yi Huang, Lihong Huang, Na Wang, Xiaofan Zhang, Shichong Zhou, Yuanyuan Wang, et al. Usfm: A universal ultrasound foundation model generalized to tasks and organs towards label efficient image analysis. *Medical image analysis*, 96:103202, 2024.
- [16] Qingbo Kang, Qicheng Lao, Jun Gao, Wuyongga Bao, Zhu He, Chenlin Du, Qiang Lu, and Kang Li. Urfm: a general ultrasound representation foundation model for advancing ultrasound image diagnosis. *iScience*, 28(8), 2025.
- [17] Sarah Leclerc, Erik Smistad, Joao Pedrosa, Andreas Østvik, Frederic Cervenansky, Florian Espinosa, Torvald Espeland, Erik Andreas Rye Berg, Pierre-Marc Jodoin, Thomas Grenier, et al. Deep learning for segmentation using an open large-scale dataset in 2d echocardiography. *IEEE transactions on medical imaging*, 38(9):2198–2210, 2019.

- [18] Jun Li, Tongkun Su, Baoliang Zhao, Faqin Lv, Qiong Wang, Nassir Navab, Ying Hu, and Zhongliang Jiang. Ultrasound report generation with cross-modality feature alignment via unsupervised guidance. *IEEE Transactions on Medical Imaging*, 2024.
- [19] Xiaodan Liang, Liang Lin, Qingxing Cao, Rui Huang, and Yongtian Wang. Recognizing focal liver lesions in ceus with dynamically trained latent structured models. *IEEE transactions on medical imaging*, 35, 10 2015.
- [20] DongAo Ma, Jiaxuan Pang, Michael B Gotway, and Jianming Liang. A fully open ai foundation model applied to chest radiography. *Nature*, pages 1–11, 2025.
- [21] Jiabo Ma, Zhengrui Guo, Fengtao Zhou, Yihui Wang, Yingxue Xu, Jinbang Li, Fang Yan, Yu Cai, Zhengjie Zhu, Cheng Jin, et al. A generalizable pathology foundation model using a unified knowledge distillation pretraining framework. *Nature Biomedical Engineering*, pages 1–20, 2025.
- [22] Xueyan Mei, Zelong Liu, Philip M Robson, Brett Marinelli, Mingqian Huang, Amish Doshi, Adam Jacobi, Chendi Cao, Katherine E Link, Thomas Yang, et al. Radimagenet: an open radiologic deep learning research dataset for effective transfer learning. *Radiology: Artificial Intelligence*, 4(5):e210315, 2022.
- [23] Adrien Meyer, Aditya Murali, Didier Mutter, and Nicolas Padoy. Ultrasam: a foundation model for ultrasound using large open-access segmentation datasets. *arXiv preprint arXiv:2411.16222*, 2024.
- [24] Lina Pedraza, Carlos Vargas, Fabián Narváez, Oscar Durán, Emma Muñoz, and Eduardo Romero. An open access thyroid ultrasound image database. In *10th International symposium on medical information processing and analysis*, volume 9287, pages 188–193. SPIE, 2015.
- [25] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.
- [26] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE transactions on pattern analysis and machine intelligence*, 39(6):1137–1149, 2016.
- [27] Xiyue Wang, Junhan Zhao, Eliana Marostica, Wei Yuan, Jietian Jin, Jiayu Zhang, Ruijiang Li, Hongping Tang, Kanran Wang, Yu Li, et al. A pathology foundation model for cancer diagnosis and prognosis prediction. *Nature*, 634(8035):970–978, 2024.
- [28] Peter NT Wells. Ultrasound imaging. *Physics in medicine & biology*, 51(13):R83, 2006.
- [29] Songxiong Wu, Hongyuan Zhang, Tingting Ye, Haoyu Xie, Ping Zeng, Qingjun Sun, Panying Wang, Bingsheng Huang, Lei Du, and Guangyao Wu. Focus: Four-chamber ultrasound image dataset for fetal cardiac biometric measurement, 2025.
- [30] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. Simmim: A simple framework for masked image modeling. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9653–9663, 2022.
- [31] Hanwen Xu, Naoto Usuyama, Jaspreet Bagga, Sheng Zhang, Rajesh Rao, Tristan Naumann, Cliff Wong, Zelalem Gero, Javier González, Yu Gu, et al. A whole-slide foundation model for digital pathology from real-world data. *Nature*, 630(8015):181–188, 2024.
- [32] Rikiya Yamashita, Tara Kapoor, Minhaj Alam, Alsfia Galimzanova, Saad Syed, Mete Akdoğan, Emel Alkim, Andrew Wentland, Nikhil Madhuripan, Daniel Goff, Victoria Barbee, Natasha Sheybani, Hersh Sagreiya, Daniel Rubin, and Terry Desser. Toward reduction in false-positive thyroid nodule biopsies with a deep learning-based risk-stratification system using us cine-clip images. *Radiology: Artificial Intelligence*, 4, 05 2022.
- [33] Siyuan Yan, Zhen Yu, Clare Primiero, Cristina Vico-Alonso, Zhonghua Wang, Litao Yang, Philipp Tschandl, Ming Hu, Lie Ju, Gin Tan, et al. A multimodal vision foundation model for clinical dermatology. *Nature Medicine*, pages 1–12, 2025.
- [34] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. In *International Conference on Learning Representations*, 2023.

- [35] Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, et al. A multimodal biomedical foundation model trained from fifteen million image–text pairs. *NEJM AI*, 2(1):A1oa2400640, 2025.
- [36] Yukun Zhou, Mark A Chia, Siegfried K Wagner, Murat S Ayhan, Dominic J Williamson, Robbert R Struyven, Timing Liu, Moucheng Xu, Mateo G Lozano, Peter Woodward-Court, et al. A foundation model for generalizable disease detection from retinal images. *Nature*, 622(7981):156–163, 2023.