TinyUSFM: Towards Compact and Efficient Ultrasound Foundation Models

Chen Ma, Jing Jiao, Shuyu Liang, Junhu Fu, Qin Wang, Zeju Li,Yuanyuan Wang, *Senior Member, IEEE*, and Yi Guo, *Member, IEEE*

***Abstract*— Foundation models for medical imaging demonstrate superior generalization capabilities across diverse anatomical structures and clinical applications. Their outstanding performance relies on substantial computational resources, limiting deployment in resource-constrained clinical environments. This paper presents TinyUSFM, the first lightweight ultrasound foundation model that maintains superior organ versatility and task adaptability of our large-scale Ultrasound Foundation Model (USFM) through knowledge distillation with strategically curated small datasets, delivering significant computational efficiency without sacrificing performance. Considering the limited capacity and representation ability of lightweight models, we propose a feature-gradient driven coreset selection strategy to curate high-quality compact training data, avoiding training degradation from low-quality redundant images. To preserve the essential spatial and frequency domain characteristics during knowledge transfer, we develop domain-separated masked image modeling assisted consistency-driven dynamic distillation. This novel framework adaptively transfers knowledge from large foundation models by leveraging teacher model consistency across different domain masks, specifically tailored for ultrasound interpretation. For evaluation, we establish the UniUS-Bench, the largest publicly available ultrasound benchmark comprising 8 classification and 10 segmentation datasets across 15 organs. Using only 200K images in distillation, TinyUSFM matches USFM's performance with just 6.36% of parameters and 6.40% of GFLOPs. TinyUSFM significantly outperforms the vanilla model by 9.45% in classification and 7.72% in segmentation, surpassing all state-of-the-art lightweight models, and achieving 84.91% average classification accuracy and 85.78% average segmentation Dice score across diverse medical devices and centers. This work successfully bridges the gap between high-performance foundation models and practical clinical deployment, winning the first place in MICCAI2025 IUGC Challenge.**

***Index Terms*— Ultrasound image analysis, Lightweight foundation model, Task adaptability**

I. Introduction

This work was supported by National Key R&D Program of China (2024YFF0507300, 2024YFF0507303), National Natural Science Foundation of China (Grant No. 62531004), and Shanghai Municipal Education Commission (Grant No. 24KNZNA09) (Corresponding authors: Yuanyuan Wang; Yi Guo).

Chen Ma, Jing Jiao, Shuyu Liang, Junhu Fu, Qin Wang, Zeju Li, Yuanyuan Wang, and Yi Guo are with the College of Biomedical Engineering, Fudan University, Shanghai 200433, China (email: cma24@m.fudan.edu.cn; jiaojing@fudan.edu.cn; syliang22@m.fudan.edu.cn; jhfu21@m.fudan.edu.cn; zejuli@fudan.edu.cn; wangqin23@m.fudan.edu.cn; yywang@fudan.edu.cn; guoyi@fudan.edu.cn).

ULTRASOUND imaging is one of the most widely used diagnostic modalities owing to its safety, affordability, and real-time capability [1]. As it plays a central role in routine clinical practice, developing reliable automated ultrasound analysis systems is crucial for improving diagnostic efficiency and reducing operator dependence. Deep learning-based approaches have achieved remarkable progress in various ultrasound tasks such as organ segmentation and disease classification [2]. However, achieving robust and generalizable automated ultrasound analysis still remains highly challenging, as image appearance is strongly influenced by operator experience, probe placement, and imaging protocol. Slight variations in probe angle or position can lead to significant structural differences even in images of the same organ. Furthermore, acquisitions from different ultrasound machines or vendors often introduce substantial discrepancies in contrast, resolution, and noise patterns. These factors result in weak generalizability of automatic analysis systems across patients, devices, and clinical settings.

Recently, medical foundation models have emerged as powerful tools for learning generalizable representations from large-scale data, demonstrating promising results across diverse clinical applications [3]. In particular, Ultrasound Foundation Model (USFM) [4] demonstrated that large-scale pre-training with spatial-frequency dual masking can produce a generalizable encoder for multiple tasks and organs. The outstanding performance of large foundation models relies on architectures with hundreds of millions of parameters, demanding significant computational and memory resources. This poses a major barrier to their practical deployment in real-world clinical settings, particularly in resource-constrained scenarios. Such practical gap motivates the development of lightweight yet high-performance ultrasound foundation models, which aim to preserve the representational capacity of large-scale foundation models while improving their feasibility for widespread clinical use [5].

Existing studies have explored model compression through architectural simplification (e.g., MobileViT [6], EfficientNet [7]) and knowledge distillation [8] from large teachers to lightweight students. The key objective is to preserve representational power and task performance while drastically reducing parameters and computation. Building lightweight ultrasound foundation models that are both efficient and effective still faces two critical challenges. First, large-scale ultrasound datasets exhibit strong heterogeneity and contain a considerable proportion of low-quality or redundant

samples [9]. Directly training lightweight models on such uncurated data can lead to underfitting or spurious feature learning [10]. Second, maintaining strong performance under strict parameter constraints requires effective knowledge transfer, yet most existing distillation methods are designed for natural images and single tasks [11], thus fail to capture the complex spectral and spatial characteristics of ultrasound [12] or build a foundation model.

To address these challenges, we propose TinyUSFM, a lightweight ultrasound foundation model that retains the versatility and adaptability of our large-scale USFM while requiring far fewer computational resources. The model employs a feature-gradient driven coreset selection strategy to curate diverse and informative samples for efficient training, and incorporates a consistency-driven dynamic distillation assisted by domain-separated masked image modeling(MIM), which leverages teacher consistency across spatial and frequency masks to guide reliable knowledge transfer while mitigating negative transfer from ambiguous samples. The main contributions of this work are summarized as follows:

- We present TinyUSFM, the first lightweight ultrasound foundation model specifically designed to maintain broad organ versatility and task adaptability of large-scale models while drastically reducing computational and memory costs. It provides a practical solution for deploying foundation-level intelligence on resource-constrained clinical platforms.
- We design a unified framework combining feature–gradient driven coreset selection with consistency-driven knowledge distillation, augmented by domain-separated masked image modeling. The former ensures diverse, high-quality, and training-beneficial data curation from large ultrasound datasets, while the latter enables reliable and domain-consistent knowledge transfer that preserves both spatial and frequency representations.
- To facilitate comprehensive and standardized evaluation, we construct UniUS-Bench, the largest publicly available ultrasound benchmark covering 15 organs with 8 classification and 10 segmentation tasks from diverse medical centers and imaging devices. Extensive experiments demonstrate that TinyUSFM outperforms all state-of-the-art lightweight models on several challenging tasks, enabling reliable and efficient clinical deployment.

II. Related Work

A. Medical Foundation Models

Foundation models have revolutionized medical AI by establishing generalist systems capable of handling diverse tasks [3]. Their success largely stems from large-scale pretraining on millions of heterogeneous images using masked image modeling or contrastive learning to capture rich semantic representations transferable across organs, modalities, and tasks [13]. Several domain-specific foundation models have been developed following this paradigm. For example, MIS-FM leverages masked autoencoding to learn universal representations from computed tomography volumes [14], RETFound employs vision transformers pretrained on millions of retinal images for fundus analysis [15], Endo-FM integrates temporal attention to capture spatial–temporal correlations in endoscopy videos [16], and MedSAM extends the general SAM architecture to medical image segmentation [17]. Most notably, Jiao et al. proposed USFM, the first universal ultrasound foundation model trained on the 3M-US database, which contains over two million ultrasound images collected from multiple organs, centers, and devices, employing spatial-frequency dual masking to address ultrasound analysis challenges [4].

Despite their strong performance, existing medical foundation models are extremely large, demanding substantial computational resources and hindering deployment in resource-constrained clinical environments [18]. This practical limitation has motivated research on lightweight foundation models that preserve generalization capacity while remaining feasible for real-world clinical use [5].

B. Knowledge Distillation

A widely adopted approach to constructing lightweight models is knowledge distillation (KD), which transfers information from high-capacity teacher models to lightweight student networks [8]. Initially applied to image classification using soft logits [19], KD has been extended to dense prediction tasks such as semantic segmentation by leveraging structured pixel-level information [20]. In medical imaging, KD has proven useful for compressing models while retaining accuracy. Qin et al. introduced an efficient segmentation architecture by distilling from well-trained networks to lightweight ones [21]. Subsequent works explored deep mutual distillation for semi-supervised segmentation to improve performance on ambiguous regions [22], multi-domain mutual distillation for enhanced representation across datasets [23], cross-layer graph flow distillation [24], and generalizable knowledge distillation [25]. While prior KD studies in medical imaging have achieved success in compressing single-task networks, extending them to foundation models is non-trivial. Foundation models must transfer broad, multi-organ and multi-task knowledge rather than task-specific features, and ultrasound data add further difficulty with strong speckle noise, shadowing, and device variability. These factors make generic distillation strategies ineffective and highlight the need for ultrasound-oriented, foundation-level approaches.

C. Coreset Selection

Foundation models are typically pretrained on massive and heterogeneous datasets that inevitably contain a considerable proportion of low-quality or redundant samples. While large teacher models possess sufficient capacity to effectively absorb useful knowledge from such uncurated data, lightweight student models with limited capacity are more susceptible to underfitting or learning spurious features. Coreset selection offers a promising solution by extracting representative and informative subsets, thereby improving both training efficiency and model robustness [26].

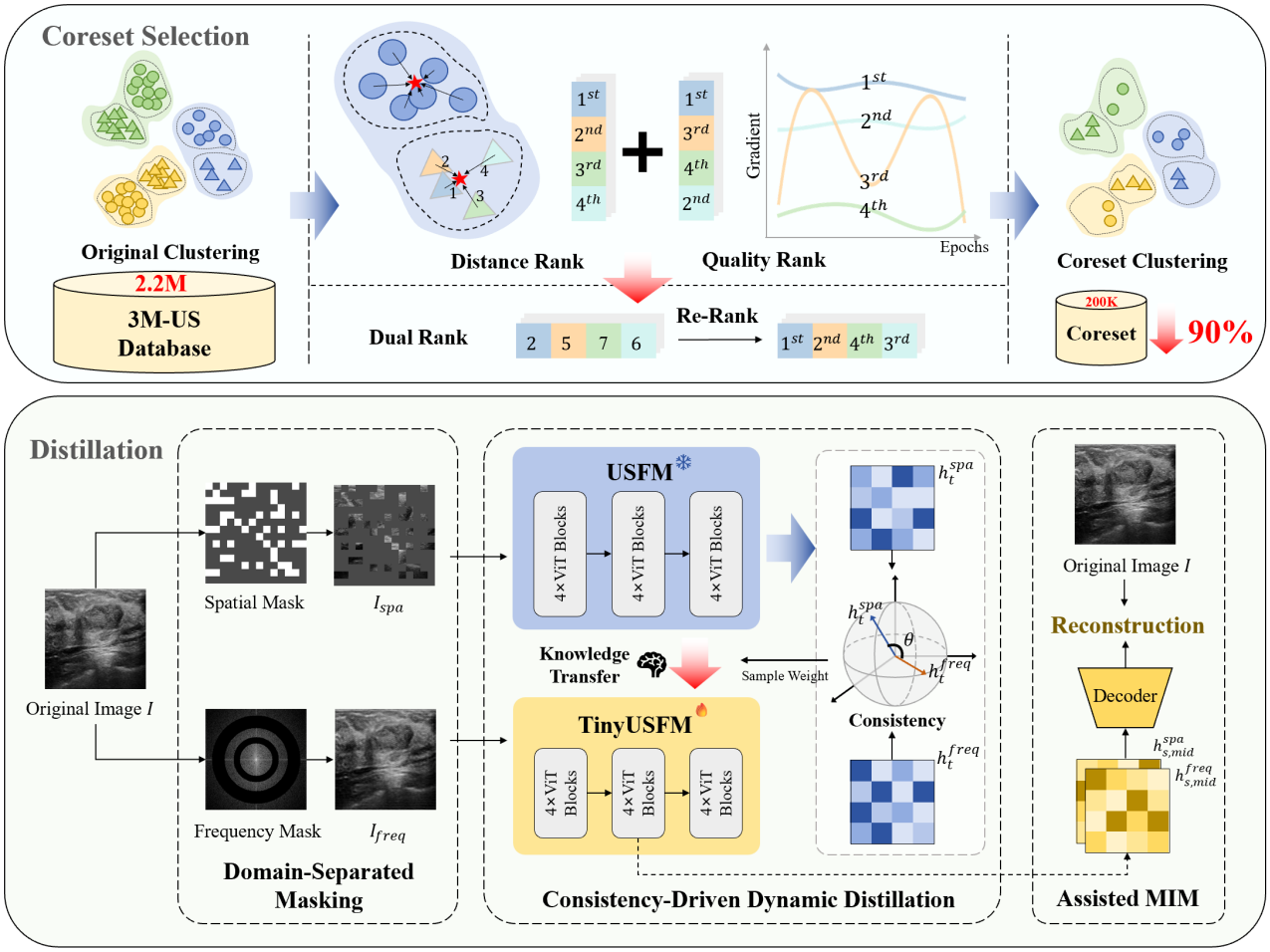


Fig. 1. Overview of proposed TinyUSFM.

Traditional approaches adopt various selection criteria, such as prototype distance [27], gradient norm [28], influence function scores [29], and uncertainty metrics [30], often combined with clustering algorithms (e.g., K-means) to mitigate redundancy [31]. However, coreset selection in medical imaging poses unique challenges. Medical datasets commonly exhibit severe class imbalance and long-tailed distributions [32], and medical foundation models are expected to capture diverse data characteristics across multiple organs and imaging devices. Existing methods, designed primarily for general-purpose vision tasks, rarely account for these domain-specific factors. Moreover, coreset selection strategies specifically tailored for foundation model construction and knowledge distillation in medical imaging remain largely unexplored.

III. Methodology

A. Method Overview

Fig. 1 illustrates the overview of our TinyUSFM framework. It transfers knowledge from the high-capacity USFM(86.5M) to lightweight TinyUSFM(5.5M) through two key components. First, a feature-gradient driven coreset selection extracts diverse and high-quality samples from USFM's pretraining 3M-US database. Second, a domain-separated MIM assisted consistency driven dynamic distillation performs deep-layer knowledge transfer combined with mid-level spatial and frequency reconstruction. The resulting TinyUSFM serves as a versatile lightweight feature encoder, enabling both efficient downstream adaptation and effective feature extraction while preserving high-quality ultrasound representations.

B. Feature-Gradient Driven Coreset Selection

Lightweight models are particularly sensitive to noisy or redundant samples, which are prevalent in large-scale heterogeneous ultrasound datasets. To ensure effective and efficient knowledge distillation under limited model capacity, we propose a feature-gradient driven coreset selection strategy that curates the most training-beneficial samples while maintaining balanced, diverse, and high-quality representation.

Our strategy jointly leverages feature diversity and training contribution, two complementary factors that capture data representativeness and learnability. This dual-drive principle combines: (1) feature-driven clustering, which hierarchically organizes the teacher model (USFM) feature space to preserve inter-organ and inter-device diversity; and (2) gradient-driven quality assessment, which quantifies each sample's learning value based on gradient stability and magnitude during teacher pretraining.

To capture multi-scale balance and diversity without over-representing dominant organs or devices, we embed all images I_i in the 3M-US database using the teacher encoder $\phi_T(\cdot)$ to obtain feature representations $z_i = \phi_T(I_i)$. A two-level hierarchical K-means structure is then applied, and both levels follow the same optimization objective function below:

$$\begin{aligned} &\min_{C,A} \sum_{i=1}^{N}\sum_{j=1}^{K} a_{i,j}\|z_i - c_j\|^2, \\ &\text{s.t. } a_{i,j} \in \{0,1\}, \sum_{j=1}^{K} a_{i,j} = 1. \end{aligned} \tag{1}$$

where $C = \{c_j\}_{j=1}^{K}$ are cluster centroids and $A = [a_{i,j}]$ is the assignment matrix. The first level partitions data by major anatomical structures, while the second refines intra-organ variations across imaging devices.

While clustering ensures diversity, it alone does not guarantee learning utility. We therefore introduce a gradient-based criterion to estimate each sample's training value based on teacher's optimization. Using gradients collected from the first ten epochs in our previous USFM pretraining, we compute mean μ_i and variance σ_i of gradients for each sample, and derive two complementary metrics: Stability score,

$$s_i^{stb} = 1 - \frac{\sigma_i}{\mu_i}, \tag{2}$$

measuring the temporal consistency, and magnitude score,

$$s_i^{mag} = \frac{\mu_i - min_j\{\mu_j\}}{max_j\{\mu_j\} - min_j\{\mu_j\}}, \tag{3}$$

representing the relative contribution to model updates. Their weighted sum

$$s_i = \alpha s_i^{stb} + (1-\alpha) s_i^{mag}, \tag{4}$$

where α is empirically set to 0.5 to balance stability and magnitude, evaluates both reliability and contribution. Samples with large and stable gradients reflect regions where the teacher learns generalizable representations, aligning data curation with the teacher's knowledge landscape.

Finally, we integrate feature diversity and gradient quality through a dual-ranking strategy. Within each fine-grained cluster, we compute the Euclidean distance from each sample to the cluster centroid. Each sample obtains two ranks, r_i^Q and r_i^D, according to its descending quality score and ascending Euclidean distance, respectively. The final selection score is calculated as

$$r_i = \beta r_i^Q + (1-\beta) r_i^D, \tag{5}$$

where β is set to 0.5 to equally weight quality and diversity. The top-k samples from each subcluster constitute the final coreset, ensuring both representational coverage and learning efficiency.

C. Domain-Separated MIM assisted Consistency-Driven Dynamic Distillation

To effectively distill knowledge from USFM to TinyUSFM, we propose a novel consistency-driven dynamic distillation framework assisted by domain-separated MIM. The core of our approach is to assess the teacher model's reliability based on its predictive consistency under spatial and frequency perturbations, thereby adaptively adjusting the distillation weight. Complementing this process, the domain-separated MIM explicitly preserves the mid-level spatial and spectral representations that are crucial for ultrasound image analysis.

1) Domain-Separated Masking Strategy: We first introduce a domain-separated masking strategy that independently perturbs ultrasound images in spatial and frequency domains to generate complementary views for both reliability assessment and representation learning. This design is motivated by the distinct roles of each domain: the spatial domain primarily captures structural and textural information, while the frequency domain encodes spectral and speckle patterns inherent to ultrasound imaging. Such a separation enables more effective evaluation of teacher model consistency and facilitates targeted, domain-specific reconstruction.

For each input ultrasound image x, the strategy generates two distinct masked views.

Spatial masking replaces randomly selected patches with the image's mean intensity:

$$M_s(I) = \begin{cases} mean(I), & (x,y) \in \textit{masked patches} \\ I(x,y), & \textit{otherwise.} \end{cases}, \tag{6}$$

where (x, y) are the coordinates in the spatial domain, and $mean(I)$ is calculated across the entire image to ensure the continuity of the grayscale distribution. This operation encourages the model to infer missing anatomical structures from contextual cues.

Frequency masking adopts a band-stop filter applied to the magnitude spectrum while preserving the phase component:

$$M_f(I) = \mathcal{F}^{-1}\big(BS(|\mathcal{F}(I)|; f_1, f_2) \cdot e^{j\phi(u,v)}\big), \tag{7}$$

where $\mathcal{F}$ and $\mathcal{F}^{-1}$ denote the Fourier and inverse Fourier transforms, $BS(\cdot)$ defines the band-stop range between f_1 and f_2, and $\phi(u, v)$ is the preserved phase. This perturbation suppresses specific frequency bands, compelling the model to recover missing spectral information.

2) Consistency-Driven Dynamic Distillation: Building upon the domain-separated masking strategy, we develop a consistency-driven dynamic distillation framework that adaptively regulates knowledge transfer, as illustrated in Fig. 2. The key idea is to evaluate the reliability of the teacher model by measuring how consistently it represents an image when different spatial or frequency masks are applied. If the deep-layer features extracted from these two masked views remain highly similar, it indicates that the teacher has captured stable and meaningful ultrasound representations. Such samples are regarded as reliable and assigned higher weights during distillation, whereas samples with inconsistent responses are down-weighted to avoid misleading supervision. This mechanism enables the student to focus on trustworthy knowledge, thereby achieving more effective and robust representation learning.

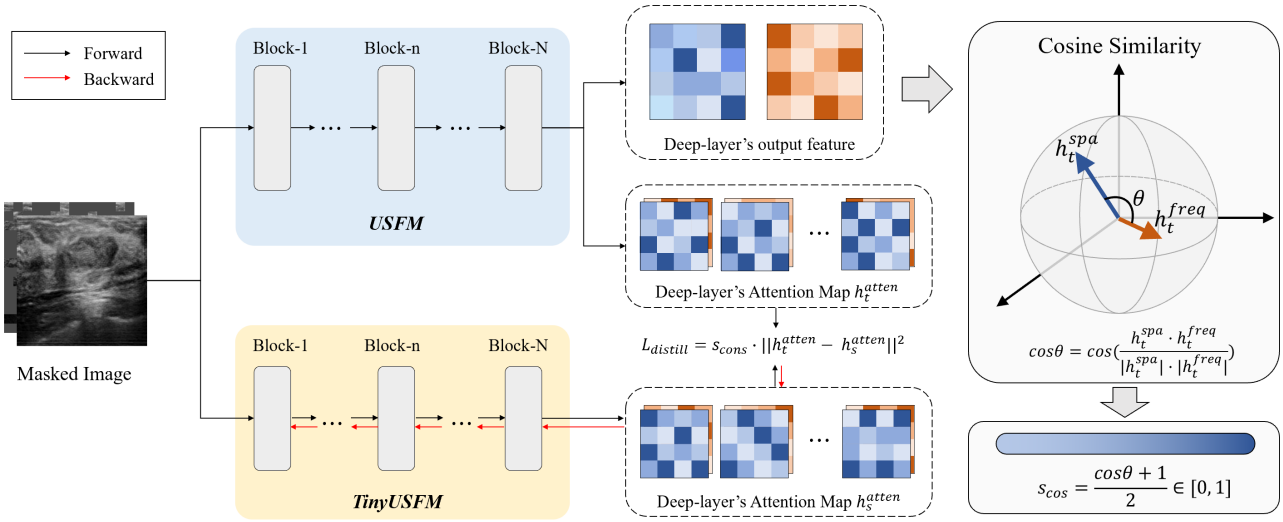


Fig. 2. Details of Consistency-Driven Dynamic Distillation.

Given the domain-separated masked inputs I_{spa} and I_{freq} obtained by the masking strategy, the teacher model $F_t(\cdot)$ extracts deep-layer representations as:

$$h_t^{spa} = F_t(I_{spa}), \quad h_t^{freq} = F_t(I_{freq}). \tag{8}$$

The teacher consistency score is computed based on cosine similarity between two representations:

$$s_{cons} = \frac{cos(h_t^{spa}, h_t^{freq}) + 1}{2}, \tag{9}$$

where $cos(\cdot)$ denotes the cosine similarity function. A higher s_{cons} indicates that the teacher encodes similar semantics under different perturbations, implying a more stable and trustworthy knowledge source.

The consistency score dynamically weights the distillation loss applied to the deep layers. The weighting is performed via attention head feature alignment, which ensures that reliable samples exert a stronger supervisory influence, while ambiguous ones are appropriately down-weighted. To enable direct head-wise feature alignment and preserve representational granularity, TinyUSFM uses a modified ViT-Tiny architecture with the same number of attention heads as the teacher USFM. This design allows one-to-one head correspondence between student and teacher for precise feature matching. Formally, the dynamic distillation loss is defined as:

$$\mathcal{L}_{distill} = s_{cons} \cdot \frac{1}{H} \sum_{h=1}^{H} \text{MSE}\left(h_s^{atten,h}, h_t^{atten,h}\right), \tag{10}$$

where $h_s^{atten,h}$ and $h_t^{atten,h}$ represent the features from the h-th attention head for student and teacher respectively.

This mechanism adaptively emphasizes reliable supervision, ensuring an efficient and stable knowledge transfer from the teacher.

3) Domain-Separated MIM: While deep-layer feature alignment effectively transfers high-level semantic knowledge, it risks causing the lightweight student to overlook crucial mid-level representations that capture fine-grained structural and spectral cues essential for ultrasound understanding. To compensate for this, we introduce a domain-separated masked image modeling (MIM) scheme that provides targeted mid-level supervision through self-reconstruction in spatial and frequency domains. By separating two reconstruction pathways, TinyUSFM can focus on learning structural textures and spectral patterns within its limited capacity, achieving stable convergence and more comprehensive knowledge transfer.

The student encoder processes both spatial and frequency masked views and extracts intermediate features h_{spa}^{mid} and h_{freq}^{mid} at the mid-layer for two complementary objectives. A lightweight decoder is then applied to recover the original inputs, which enables the student to learn informative representations through self-supervised reconstruction. Specifically, the reconstructed outputs are obtained as follows:

$$\hat{I}_{spa} = F_D(h_{spa}^{mid}), \quad \hat{I}_{freq} = F_D(h_{freq}^{mid}). \tag{11}$$

The combined objective for mid-level layer reconstruction is then defined as:

$$\mathcal{L}_{recon} = \text{MSE}(\hat{I}_{spa}, I) + \text{MSE}(\hat{I}_{freq}, I), \tag{12}$$

where $\lambda = 1.0$ balances the contributions from deep-layer dynamic distillation and mid-level reconstruction assistance.

This integrated framework enables TinyUSFM to efficiently acquire both spatial and spectral representations while maintaining computational efficiency, allowing the student model to effectively absorb the teacher's knowledge and achieve robust, high-fidelity representations for ultrasound image analysis.

IV. Experiments

A. Datasets

Our study utilizes two purpose-specific datasets. For pre-training, we apply our feature-gradient driven coreset selection strategy to curate a 200K subset from the large-scale 3M-US database [4] to enable efficient knowledge distillation.

For evaluation, we introduce UniUS-Bench, the largest publicly available ultrasound benchmark designed for standardized and comprehensive assessment of ultrasound foundation models. Unlike previous datasets that are often limited to a

TABLE I
SUMMARY OF THE UNIUS-BENCH, INCLUDING 8 CLASSIFICATION AND 10 SEGMENTATION DATASETS ACROSS 15 ORGANS.

Downstream	Category	Dataset	Organs	#Images	Targets
Classification	Binary	CUBS [33]	Carotid	1378	Non-FUP / FUP Event
		UF1990 [34]	Uterine	1990	Normal / Fibroid
		TN3K [35]	Thyroid	3491	Benign / Malignant
		STMUS [36]	Skeletal muscle	5312	Normal / Pathological
	Multi-class	AUL [37]	Liver	735	Normal / Benign / Malignant
		BUSI [38]	Breast	780	Normal / Benign / Malignant
		MMOTU [39]	Ovarian	1469	8 Subtype tumors
		Fetal Planes [40]	Fetus	12400	6 Fetal planes
Segmentation	Single-target	Luminous [41]	Multifidus muscle	341	Lumbar multifidus muscle
		KidneyUS [42]	Kidney	487	Kidney capsule
		GIST514 [43]	Stomach	514	Gastrointestinal stromal tumor
		DDTI [44]	Thyroid	637	Thyroid nodule
		MMOTU [39]	Ovarian	1469	Ovarian tumor
		BUSBRA [45]	Breast	1875	Breast tumor
		Ultrasound Nerve Seg. [46]	Neck Nerve	5735	Brachial Plexus
	Multi-target	LUSS [47]	Lung	564	Rib / Pleural line / A-line / B-line / B-line confluence
		FH-PS-AoP [48]	Pelvis	4000	Pubic symphysis / Fetal head
		CAMUS [49]	Cardiac	19232	LV cavity / LV myocardium / LA

single organ or task, UniUS-Bench integrates diverse clinical scenarios from different medical centers and imaging devices into a unified evaluation suite. It encompasses 8 classification and 10 segmentation datasets, covering 15 organs with a total of 60,940 ultrasound images and 62,409 annotations. Importantly, UniUS-Bench is built entirely from existing public datasets without including any private data, ensuring openness, transparency, and reproducibility, thereby making it the most extensive and clinically relevant benchmark to date.

The benchmark includes diagnostic classification tasks covering both binary and multi-class organ-level diagnosis (e.g., liver, breast, fetal planes), as well as segmentation tasks spanning single-target and multi-target structures at both organ-level (e.g., kidney, fetal head) and tumor-level (e.g., breast tumor, thyroid nodule, liver lesion). This comprehensive design enables systematic evaluation of model generalization across diverse anatomical structures, imaging devices, and clinical tasks.

Detailed data statistics, including sample size, task type, and organ coverage, are summarized in Table I. To enable fair comparison, we follow the official training, validation and testing splits whenever available. For those datasets without predefined splits, we adopt a stratified random partition with a ratio of 7:1:2 for training, validation, and testing, respectively.

B. Compared methods

To comprehensively evaluate TinyUSFM, we compare it with three categories of representative methods, including conventional models, lightweight models, and foundation models.

For the classification task, we consider the following methods: (1) conventional CNN/Transformer models: ResNet [50] and ViT-B [51], (2) lightweight models specifically designed for efficiency: MobileViT [6], EfficientNet [7], and ViT-Tiny [52], and (3) foundation models: SimMIM [53] and our teacher model USFM [4]. Model performance is reported in terms of classification accuracy.

For the segmentation task, we evaluate TinyUSFM against the following methods: (1) conventional segmentation models: TransUNet [54] and SwinUNet [55]; (2) lightweight segmentation models: SegFormer [56], SeaFormer [57], and ViT-Tiny with FPN decoders, specifically designed for efficiency; and (3) foundation models: SimMIM and the teacher model USFM. Model performance is reported in terms of Dice score.

C. Implementation details

All input images were resized to 224×224. During pretraining, the spatial masking ratio was set to 0.75. For frequency masking, we followed the USFM configuration by defining seven band-pass filters spanning from low to high frequencies and randomly selecting two filters to generate the mask. A 10×10 area at the center of the frequency spectrum, containing crucial low-frequency information, was always preserved. The frequency masking ratio was set to 0.4.

For downstream tasks, training employed a warmup poly learning rate scheduler with an initial learning rate of 1e-4 for 400 epochs. Data augmentation strategies were tailored to each task. For classification, we employed random flipping, rotation, and scaling. For segmentation, we additionally employed contrast and gamma adjustments to enhance model robustness.

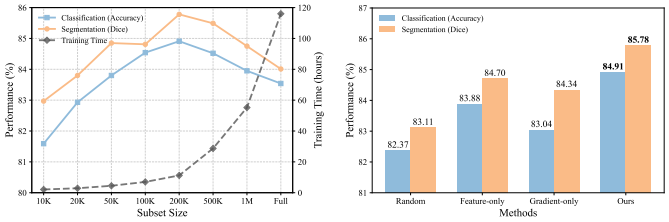
All experiments were conducted on an AMD EPYC 7763 CPU and NVIDIA A100 GPU. All models were developed using PyTorch. All compared methods were implemented strictly following their official repositories or the settings reported in their original papers to ensure fair comparison.

V. RESULTS

We present a comprehensive evaluation of TinyUSFM, assessing its pre-training efficacy and downstream adaptability.

TABLE II
OVERALL FRAMEWORK ABLATION STUDY. RESULTS ARE REPORTED AS CLASSIFICATION ACCURACY (%) AND SEGMENTATION DICE SCORE (%) ON UNIUS-BENCH. BEST RESULTS ARE IN **BOLD**.

Components			Classification	Segmentation
Coreset	MIM	Distillation		
×	×	×	75.46	78.06
×	×	✓	81.93	82.03
×	✓	✓	83.54	84.01
✓	×	✓	83.50	84.59
✓	✓	✓	**84.91**	**85.78**



(a) Ablation on subset size (b) Comparison at 200K subset

Fig. 3. Ablation study on coreset selection strategy.

The evaluation first analyzes how the distillation framework enhances representation quality and efficiency. It then benchmarks TinyUSFM against conventional, lightweight, and foundational models across diverse tasks. Together, these experiments provide a holistic evaluation of how TinyUSFM achieves efficient and reliable knowledge transfer while maintaining lightweight deployment.

A. Effectiveness of TinyUSFM pretraining

To comprehensively validate the effectiveness of TinyUSFM pretraining, we conduct systematic analyses from multiple perspectives. Specifically, we first perform ablation studies on the overall framework to assess the contribution of each proposed component, followed by targeted studies on coreset selection efficiency and domain-separated reconstruction superiority. We further employ UMAP [58] for representation visualization to examine TinyUSFM's discriminability and generalization across organs and tasks. Finally, we compare the model size and computational cost to highlight the compactness and efficiency of our TinyUSFM.

Table II presents comprehensive ablation results evaluating the contribution of each component: feature-gradient driven coreset selection, domain-separated MIM, and consistency-driven dynamic distillation. The baseline configuration without any components achieves 75.46% average classification accuracy and 77.97% average segmentation Dice score on UniUS-Bench. Adding consistency-driven dynamic distillation alone provides substantial improvements to 81.93% and 82.03% respectively, demonstrating the effectiveness of dynamic distillation weighting and the strong representational capability of USFM. With the addition of domain-separated MIM as an auxiliary task, distillation achieves enhanced performance of 83.54% and 84.01%, highlighting the importance of mid-level reconstruction learning. Incorporating feature-gradient coreset selection with distillation yields 83.50% accuracy and 84.59% Dice score, demonstrating that selecting high quality and non-redundant data benefits model distillation by reducing noisy supervision. The complete framework integrating all three components achieves optimal performance with 84.91% classification accuracy and 85.78% segmentation Dice score, validating the synergistic effect of combining these three components.

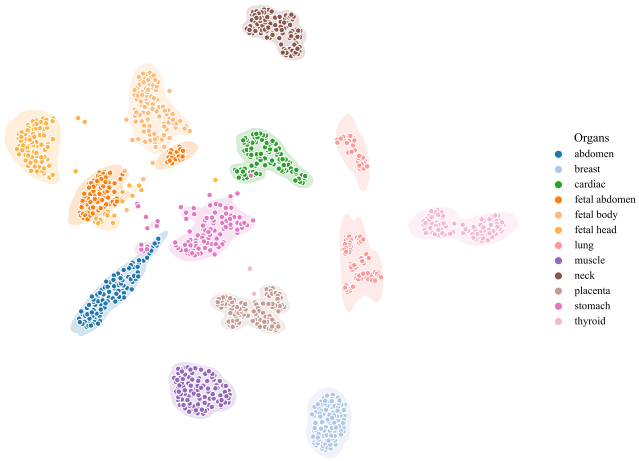
1) Efficiency of coreset selection: Fig. 3 illustrates the ablation study conducted to evaluate our coreset selection strategy. We investigate the impact of subset size by constructing training sets ranging from 10K to the full 3M-US dataset containing over 2.1 million images. This experiment investigates the relationship between dataset size and its effects on the performance of lightweight models and the efficiency of distillation. In addition, after determining 200K as the optimal subset size, we further compare our feature–gradient driven selection against several alternative strategies, including random sampling, feature-only clustering, and gradient-only quality filtering. This design allows us to assess both the optimal subset size and the relative advantage of combining diversity preservation with quality-based ranking.

The results reveal two key observations. First, performance consistently improves as the subset size grows from 10K to 200K, with average classification accuracy rising from 81.59% to 84.91% and average segmentation Dice score from 82.97% to 85.78%. Notably, the optimal 200K subset, using merely 10% of the full dataset, outperforms training on the complete 3M-US database by 1.37% in classification and 1.77% in segmentation, while requiring less than 10% of the training time on a single NVIDIA A100 GPU. This finding demonstrates that curated high-quality data can be more effective and efficient than large redundant datasets for lightweight models. Beyond 200K, both accuracy and Dice score begin to decline (83.54% and 84.01% at full dataset) and the training time increases substantially, suggesting that additional low-quality samples not only compromise representation quality but also introduce unnecessary computational overhead.

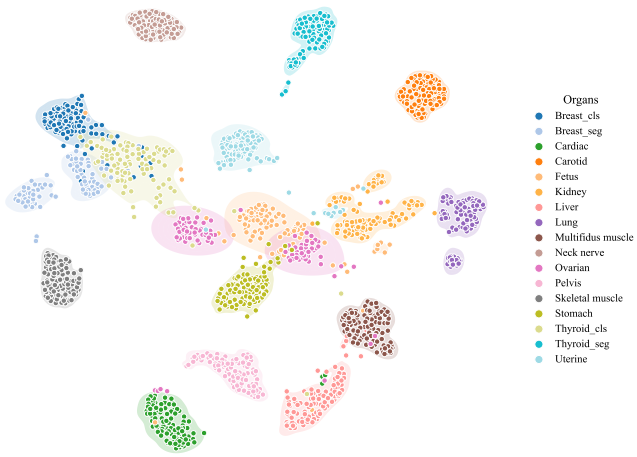
Second, our integrated approach significantly outperforms all baselines at the 200K subset: achieving +2.54% in classification accuracy and +2.67% in segmentation Dice over random sampling, +1.03% and +1.08% over feature-only clustering, and +1.87% and +1.44% over gradient-only filtering, respectively. These findings underscore that effective lightweight model training relies not merely on data reduction, but on carefully curated data that balances diversity with quality, thereby validating the design of our coreset selection method.

2) Superiority of Domain-Separated Reconstruction: To further validate the superiority of the proposed domain-separated MIM, we conduct comprehensive analyses of the best reconstruction layer and the reconstruction domain configuration.

Fig. 5a investigates the optimal layer for applying domain-separated MIM reconstruction in our distillation framework. We evaluate reconstruction at different layers while keeping other components fixed. Both classification and segmentation tasks show consistent trends, with performance gradually



(a) Distribution of 3M-US Database in TinyUSFM feature space.



(b) Distribution of UniUS-Bench in TinyUSFM feature space.

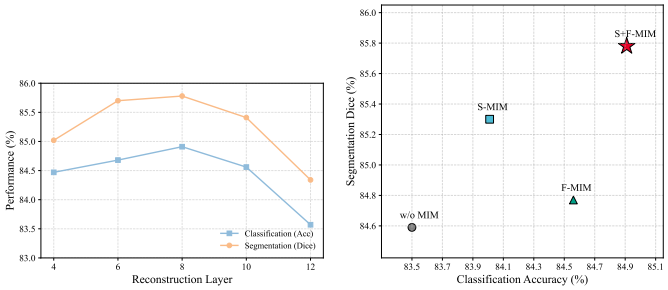
Fig. 4. UMAP visualization of TinyUSFM feature representations.

improving from shallow layers to mid-level layers, peaking at the 8th layer with 84.91% average classification accuracy and 85.78% average segmentation Dice score. Notably, performance significantly degrades when applying reconstruction at the deep-layer for both tasks, dropping to 83.57% in classification and 84.34% in segmentation. This suggests that performing reconstruction and deep-layer feature alignment simultaneously at the same layer creates conflicting objectives that impair learning effectiveness. These findings validate our design strategy of placing reconstruction at the mid-layer and feature alignment at the deep-layer, which achieves optimal performance while avoiding optimization conflicts between different learning objectives.

In addition, we examine the effect of different reconstruction domains, including spatial-only (S-MIM), frequency-only (F-MIM), and spatial–frequency combined (S+F-MIM) configurations, alongside the baseline without MIM. As illustrated in Fig. 5b, the baseline without MIM achieves 83.50% classification accuracy and 84.59% segmentation Dice score. Incorporating spatial reconstruction (S-MIM) raises performance by +0.51% and +0.71%, while frequency reconstruction (F-MIM) further improves it by +1.06% and +0.18%, confirming that both domains contribute complementary information beneficial for ultrasound representation learning. The domain-separated MIM achieves the highest performance, significantly outperforming both single-domain variants. This superiority arises because spatial reconstruction enhances structural and texture perception, whereas frequency reconstruction captures spectral and noise-related characteristics unique to ultrasound imaging.

Together, these findings demonstrate that performing mid-level reconstruction with domain-separated MIM offers the most effective and stable configuration. This design allows the lightweight TinyUSFM to capture comprehensive spatial–spectral representations, thereby improving the deep-layer knowledge transfer quality and overall model robustness.

3) Discriminability and Generalizability of Learned Feature Representations: To qualitatively analyze the feature representations learned during TinyUSFM pretraining, we randomly sample 100 images per organ from both the curated 3M-US pretraining subset and the UniUS-Bench. The embeddings of these samples are visualized using UMAP, enabling inspection of organ-level distribution and similarity patterns in the learned latent space. By projecting high-dimensional features to two dimensions, we assess how well the pretraining captures inter-organ separability and whether the representations generalize across downstream classification and segmentation tasks.



(a) Performance variation across different reconstruction layers.

(b) Comparison of reconstruction domain configurations.

Fig. 5. Ablation studies on domain-separated reconstruction.

Fig. 4 demonstrates that TinyUSFM successfully learns organ-specific representations during pretraining. In the 3M-US feature space (Fig. 4a), different organs form distinct, well-separated clusters with clear boundaries, indicating that the model captures meaningful anatomical differences across ultrasound imaging of various body parts, thereby exhibiting strong discriminability of organ-level representations. Each organ type occupies a distinct region in the feature space, demonstrating effective discrimination between different anatomical structures.

The distribution of UniUS-Bench in TinyUSFM feature space (Fig. 4b) reveals strong generalization capability from pretraining to downstream tasks. Despite training solely on the 3M-US subset, TinyUSFM maintains clear feature separability

TABLE III
CLASSIFICATION PERFORMANCE COMPARISON ON UNIUS-BENCH. RESULTS ARE REPORTED AS ACCURACY (%). BEST RESULTS ARE IN **BOLD-UNDERLINED**, SECOND BEST ARE IN **BOLD**.

Types	Conventional Model		Lightweight Model			Foundation Model		
Models	ResNet50	ViT-B	EfficientNet	MobileViT	ViT-T	SimMIM	USFM	TinyUSFM
#Params	23.5M	86.5M	7.7M	4.4M	5.5M	86.5M	86.5M	5.5M
GFLOPs	8.21	33.72	1.34	2.86	2.16	33.72	33.72	2.16
Carotid	77.54±1.54	81.16±1.24	78.99±0.89	81.88±0.96	81.88±0.96	81.16±0.74	**86.23±0.40**	**84.32±0.67**
Uterine	91.67±0.28	89.39±0.63	93.43±0.27	95.20±0.28	90.66±0.28	94.70±0.25	**96.72±0.22**	**96.21±0.19**
Thyroid	64.27±1.31	65.91±1.23	67.05±1.33	64.11±1.09	59.05±2.51	69.66±1.30	**72.59±0.67**	**71.45±0.71**
Muscle	87.18±0.41	86.42±0.59	89.36±0.45	89.17±0.47	88.32±0.49	91.32±0.31	**92.59±0.29**	**91.93±0.28**
Liver	74.32±0.89	70.27±1.16	74.32±0.93	72.30±0.95	68.92±1.20	75.68±0.89	**78.38±0.75**	**77.70±0.83**
Breast	83.33±1.30	79.17±1.77	86.67±1.32	84.17±1.29	70.83±2.38	85.00±1.23	**88.33±1.18**	**87.50±1.25**
Ovarian	66.10±1.53	60.98±2.58	67.38±1.53	69.30±1.61	57.57±2.91	70.59±1.28	**75.48±1.56**	**76.33±1.30**
Fetus	83.60±0.43	86.25±0.27	91.96±0.11	92.05±0.18	86.47±0.23	93.05±0.19	**93.74±0.13**	**93.83±0.10**
Avg	78.50	77.44	81.14	81.02	75.46	82.42	**85.51**	**84.91**

across diverse downstream datasets, even for organs previously unseen during pretraining. Notably, related anatomical structures exhibit spatial proximity in the embedding space (e.g., breast classification and segmentation tasks cluster closely), while visually similar organs remain well separated (e.g., ovarian and uterine tasks show distinct boundaries). This demonstrates that TinyUSFM learns transferable ultrasound-specific representations with strong discriminability, enabling effective generalization across diverse datasets, imaging protocols, and clinical applications. These results validate the effectiveness of our novel knowledge distillation framework for lightweight ultrasound foundation model construction and highlight its potential to serve as a universal representation backbone for a wide range of downstream tasks.

4) Deployability of TinyUSFM: To evaluate the practical deployability of TinyUSFM, we analyze its computational efficiency, memory consumption, and inference speed. Benefiting from its compact architecture and efficient distillation framework, TinyUSFM contains only 5.5M parameters and requires 2.16 GFLOPs, corresponding to just 6.36% and 6.40% of USFM's parameters and computation, respectively. When deployed on a single GPU, TinyUSFM demonstrates outstanding efficiency. On an RTX 2080 Ti, using a batch size of 16, the average inference time per image is below 1 ms, and the total GPU memory usage remains under 1 GB. This indicates that the model can run smoothly even on devices with very limited computational resources, without the need for high-end accelerators or large memory. Despite this extreme compactness, TinyUSFM maintains nearly identical performance to its large-scale teacher USFM.

B. Downstream tasks adaptation

We conduct systematic evaluations of TinyUSFM on UniUS-Bench for two downstream tasks: classification and segmentation.

1) Diagnostic Classification across Organs: TinyUSFM's features are fed into a linear layer to predict target labels. This simple design ensures efficient downstream adaptation while fully leveraging TinyUSFM's representations.

Table III presents comprehensive classification results across the 8 datasets in UniUS-Bench, comparing TinyUSFM against three categories of baseline methods.

TinyUSFM significantly outperforms conventional architectures including ResNet50(78.50%) and ViT-B(77.44%), while using minimal parameters and computational resources, demonstrating the effectiveness of foundation model pretraining for ultrasound representation learning. Among lightweight architectures, TinyUSFM demonstrates competitive performance against EfficientNet(81.14%) and MobileViT(81.02%). In addition, it improves the vanilla ViT-Tiny's performance from 75.46% to 84.91%, representing a 9.45% improvement. Notably, TinyUSFM achieves remarkable breakthroughs on challenging tasks compared to ViT-Tiny, such as ovarian tumor classification (76.33% vs 57.57%), breast tumor diagnosis (87.50% vs 70.83%), and liver lesion detection (77.70% vs 68.92%). While these lightweight models are specifically designed for computational efficiency, TinyUSFM leverages ultrasound domain knowledge through our distillation framework, highlighting the importance of domain-specific adaptation beyond generic efficiency-oriented designs.

Most importantly, TinyUSFM approaches the performance of its teacher model USFM (84.91% vs 85.51%), with only a 0.60% performance gap while achieving 15.7× parameter reduction and 15.6× computational efficiency improvement. The knowledge distillation proves particularly effective on complex diagnostic tasks, even surpassing the teacher model performance on fetal plane classification (93.83% vs 93.74%) and ovarian tumor subtype classification (76.33% vs 75.48%). This near-optimal knowledge transfer confirms the effectiveness of our proposed distillation framework in compressing ultrasound foundation models without sacrificing task performance.

Additionally, the performance gains are consistent across diverse ultrasound imaging scenarios. TinyUSFM shows particularly strong performance on challenging multi-class tasks including fetal plane (93.83%), which distinguishes six standard fetal views, and breast tumor (87.50%), which differentiates normal tissue, benign masses, and malignant lesions. The

TABLE IV
SEGMENTATION PERFORMANCE COMPARISON ON UNIUS-BENCH. RESULTS ARE REPORTED AS DICE SCORE (%). BEST RESULTS ARE IN **BOLD-UNDERLINED**, SECOND BEST ARE IN **BOLD**.

Type	Conventional Model		Lightweight Model			Foundation Model		
Model	TransUNet	SwinUNet	SegFormer	SeaFormer	ViT-T	SimMIM	USFM	TinyUSFM
#Params	105.3M	27.2M	13.7M	13.9M	6.2M	101.2M	101.2M	6.2M
GFLOPs	58.46	11.79	5.07	2.53	3.45	53.07	53.07	3.45
Muscle	90.06±9.19	86.09±10.18	90.24±8.70	90.49±8.81	88.41±9.60	89.45±9.26	**92.04±7.29**	**91.92±6.91**
Kidney	93.27±4.14	87.91±8.42	89.98±6.91	91.98±4.88	87.76±8.91	89.23±7.02	**93.82±4.19**	**93.46±3.97**
Stomach	72.19±28.65	60.61±30.06	76.00±25.35	73.67±27.27	61.43±29.01	73.16±27.36	**81.87±19.22**	**80.30±21.58**
Thyroid	**83.36±14.46**	77.64±20.57	80.34±19.35	80.06±19.22	70.52±21.01	83.29±13.88	**83.82±14.25**	82.77±16.37
Ovarian	83.21±20.38	81.41±22.80	79.17±24.63	81.03±21.12	76.91±21.85	**83.49±21.57**	**84.56±19.98**	82.65±21.91
Breast	87.89±14.53	80.33±19.45	86.17±12.25	86.94±13.80	73.71±23.26	85.69±13.86	**90.10±9.97**	**88.88±11.05**
Neck	78.02±15.48	77.84±15.49	73.65±14.90	73.78±14.90	72.22±17.99	77.05±17.01	**79.25±15.28**	**80.00±14.31**
Lung	**71.26±30.63**	64.16±33.39	68.00±31.75	66.77±32.34	64.83±33.33	66.30±32.62	**71.35±30.77**	69.57±30.91
Pelvis	96.56±2.89	95.52±3.49	96.13±3.01	95.85±3.14	96.02±3.14	96.11±3.16	**97.36±2.70**	**97.44±2.63**
Cardiac	90.32±4.88	89.74±5.63	90.17±5.10	90.13±5.12	88.77±6.23	89.21±5.83	**90.46±4.98**	**90.84±4.80**
Avg	84.61	80.13	82.99	83.07	78.06	83.30	**86.46**	**85.78**

model achieves its most significant improvement on ovarian tumor subtype classification (76.33%), an eight-class task involving various histopathological types, which is notoriously difficult due to subtle intra-class variations, highly similar texture patterns among subtypes, and substantial class imbalance. For binary classification tasks, TinyUSFM maintains robust performance across different anatomical regions, including Carotid artery (84.32%) and Liver lesion (77.70%). The consistent performance across organs and task complexity validates the generalizability of our learned representations and the effectiveness of our feature-gradient driven coreset selection in capturing diverse ultrasound imaging patterns.

2) Anatomical Segmentation across Structures: In the segmentation task, TinyUSFM serves as the backbone to extract multi-scale feature maps from intermediate layers. The extracted features are subsequently processed by a lightweight pyramid neck and decoded through an FPN-style head to produce dense prediction maps. Hierarchical features are extracted from layers 3, 5, 7, 11, and the fused representation is upsampled to match the input resolution. This design maintains a compact architecture while enabling accurate pixel-wise predictions.

Table IV summarizes segmentation performance across ten datasets in UniUS-Bench.

TinyUSFM consistently outperforms both conventional segmentation models and lightweight segmentation models. Compared to conventional models, our approach surpasses TransUNet (84.61%) and SwinUNet (80.13%) by margins of 1.17%, and 5.65% respectively. Against lightweight architectures, TinyUSFM shows improvements of 2.79% over SegFormer (82.99%) and 2.71% over SeaFormer (83.07%), while using fewer parameters and comparable computational resources. Most notably, TinyUSFM significantly outperforms the baseline ViT-Tiny backbone by 7.72% (85.78% vs 78.06%), achieving remarkable improvements on challenging segmentation tasks like gastrointestinal stromal tumor segmentation (80.30% vs 61.43%), and breast tumor segmentation (88.88% vs 73.71%), validating our ultrasound-specific distillation framework. Furthermore, TinyUSFM approaches the performance of its teacher model USFM (85.78% vs 86.46%) with only a 0.68% gap while achieving 16.3× parameter reduction and 15.4× computational efficiency improvement in segmentation. Remarkably, TinyUSFM even surpasses its teacher model on specific challenging segmentation tasks like neck nerve (80.00% vs 79.25%) and cardiac chamber (90.84% vs 90.46%), demonstrating that the MIM assisted distillation process can enhance performance on anatomically complex structures.

The performance gains are consistent across diverse anatomical structures and pathological conditions. TinyUSFM excels in challenging tumor segmentation tasks such as breast tumor (88.88%), and thyroid nodule (82.77%). For organ segmentation tasks, robust performance is maintained across organs including kidney (93.46%), muscle (91.92%), and cardiac (90.84%).

To provide a visual comparison, Fig. 6 presents segmentation results across ten representative organs from UniUS-Bench. Compared with conventional (TransUNet, SwinUNet) and lightweight (SegFormer, SeaFormer) models, TinyUSFM produces more accurate and smoother contours that better align with anatomical boundaries, particularly in structurally complex regions such as the cardiac chambers and neck nerve. The baseline ViT-Tiny exhibits inconsistent delineations and boundary fragmentation, indicating insufficient ultrasound representation learning. By contrast, TinyUSFM maintains high spatial coherence and captures subtle tissue interfaces even under strong speckle noise or low-contrast conditions. The performance advantage is especially evident for tumor-level segmentation (e.g., breast and ovarian lesions), where TinyUSFM yields precise shape preservation and reduced false positives. These qualitative results further confirm that our domain-separated distillation effectively enhances spatial–spectral representation learning and yields anatomically reliable predictions across diverse ultrasound organs and ac-

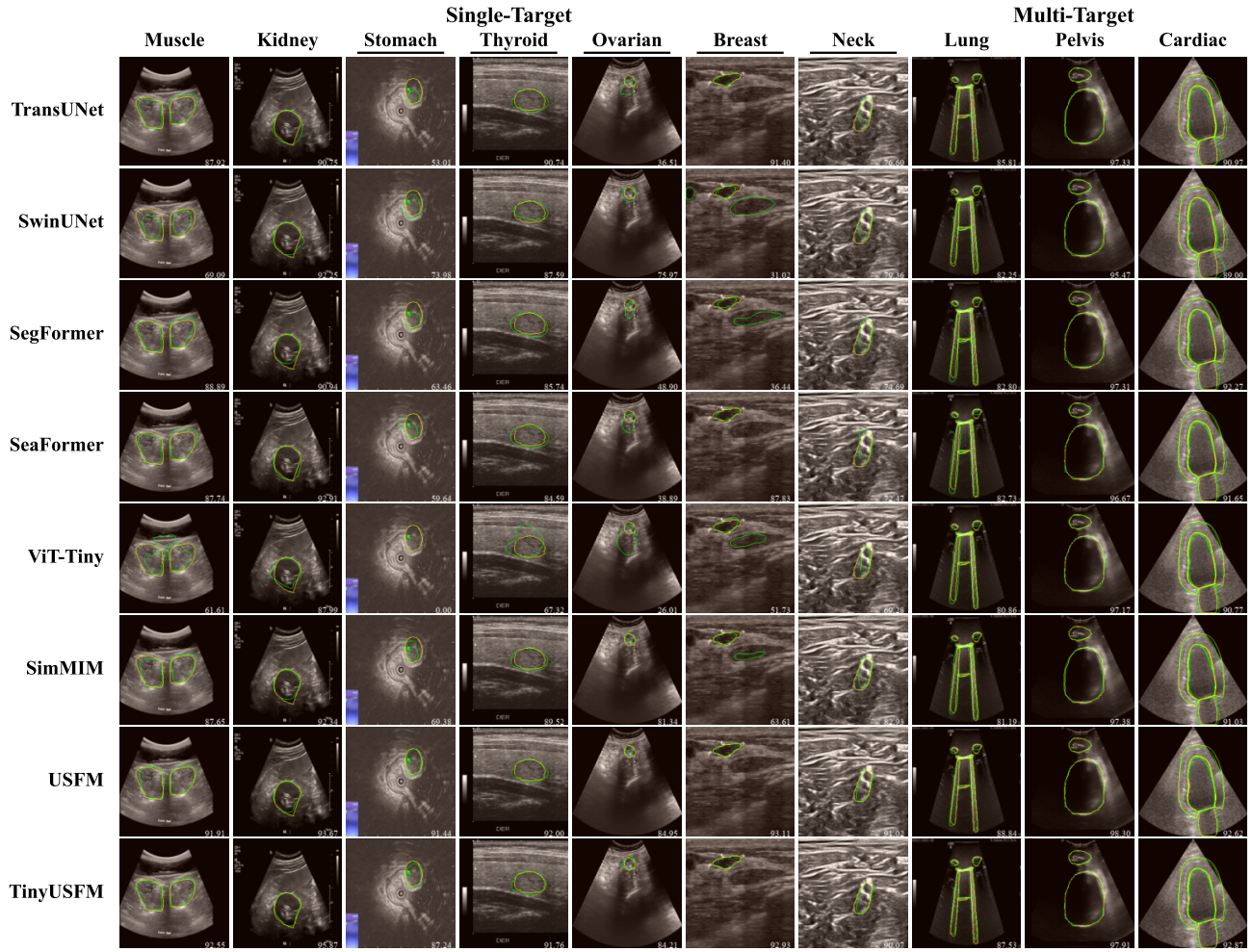


Fig. 6. Qualitative comparison of segmentation results across different organs on UniUS-Bench. Depicted in yellow is the ground truth, and in green is the predicted result. Tumor-level datasets are indicated by underlined titles.

quisition conditions.

These results demonstrate that our framework effectively transfers ultrasound-specific knowledge to lightweight architectures, surpassing all lightweight models and matching state-of-the-art performance across both classification and segmentation tasks while enabling practical clinical deployment through significantly reduced computational requirements.

VI. Discussion

The proposed TinyUSFM pioneers a teacher-guided approach for building lightweight foundations in ultrasound imaging. Instead of pursuing architectural compression, it emphasizes data- and representation-centric efficiency through effective knowledge transfer from large teacher model to compact student. This teacher-guided paradigm enables TinyUSFM to retain high representational fidelity with minimal computational cost. The following discussion analyzes how this design narrows the gap with large foundation models, surpasses conventional lightweight networks, and facilitates clinical deployment, followed by limitations and future directions.

A. Tiny model vs large scale foundation model

A key finding of this study is that carefully distilled lightweight models can approach the performance of large-scale foundation models without inheriting their prohibitive computational demands. Large ultrasound foundation models, such as USFM, rely on hundreds of millions of parameters and dozens of GFLOPs, which makes them unsuitable for deployment in most clinical environments. By contrast, TinyUSFM achieves comparable accuracy and segmentation quality with only 5.5M parameters and 2.16 GFLOPs which is over an order-of-magnitude reduction.

This performance is not a trivial result of model compression, but instead reflects the importance of data curation. Our feature-gradient driven coreset selection demonstrates that a subset of only 200K well-chosen samples can provide richer training signals for lightweight models than the entire 3M-US dataset. This result suggests that lightweight architectures benefit more from quality-focused sampling than from sheer data volume, which often introduces noise and redundancy. The observation that the curated subset even outperforms full-dataset training indicates that lightweight models have different optimization dynamics compared with large-scale

networks: whereas large models can tolerate noisy gradients, small models require stable, representative, and information-rich supervision. This insight highlights a paradigm shift for building tiny foundation models, emphasizing data efficiency and representational fidelity rather than brute-force scale.

B. TinyUSFM vs lightweight specific model

Conventional lightweight models, such as MobileViT and EfficientNet, achieve efficiency by architectural simplification or parameter reduction, but they typically lack adaptability across heterogeneous medical tasks. Their design philosophy emphasizes general-purpose efficiency rather than domain-specific knowledge transfer, leading to limited gains in complex clinical scenarios. TinyUSFM, however, is not just another small network. Its advantage comes from an ultrasound-oriented distillation strategy that explicitly addresses the spectral–spatial complexity of ultrasound signals.

The consistency-driven dynamic distillation selectively emphasizes teacher knowledge that is stable under domain-separated perturbations, preventing the student from inheriting unreliable supervision. Meanwhile, the auxiliary domain-separated MIM reconstruction enables the student to retain mid-level structural and spectral cues that would otherwise be lost during aggressive compression. Ablation studies confirm the complementarity of these components, but more importantly, their integration demonstrates that domain-aware distillation can outperform architecture-centric efficiency. This explains why TinyUSFM not only surpasses vanilla ViT-Tiny by large margins (+9.45% in classification and +7.72% in segmentation), but also outperforms efficiency-oriented models designed from scratch. In essence, TinyUSFM illustrates a new design pathway for lightweight foundation models: knowledge transfer tailored to domain characteristics rather than solely architectural reduction.

C. Clinical significance

The clinical relevance of TinyUSFM lies in bridging the gap between cutting-edge foundation models and real-world deployment constraints. Large-scale models often require specialized hardware, which is inaccessible in most community hospitals and point-of-care settings. By drastically reducing model size and computational demand without compromising diagnostic accuracy, TinyUSFM can be integrated into portable ultrasound scanners, mobile workstations, and edge devices. This not only lowers the barrier to adoption but also democratizes access to advanced AI support in resource-limited regions. Another dimension of clinical significance is workflow integration. A single lightweight model that supports both diagnostic classification and precise anatomical segmentation reduces the need for maintaining multiple specialized networks. This simplification improves reliability, shortens inference time, and lowers technical complexity for clinicians. Beyond efficiency, TinyUSFM has demonstrated robustness in external validation: winning first place in the MICCAI2025 IUGC Challenge, where it generalized successfully to unseen datasets. Such independent validation highlights the model's scalability and reinforces confidence in its clinical readiness. Collectively, these features suggest that TinyUSFM is not merely an academic exercise in model compression, but a practical enabler of foundation-level intelligence in everyday clinical practice.

D. Limitations and future directions

This work primarily focuses on two-dimensional ultrasound imaging, leaving modalities such as 3D ultrasound, Doppler, or elastography for future exploration. Expanding to these modalities, and to additional tasks like lesion detection or image enhancement, will further test the generalizability of the proposed framework. Nonetheless, the present results already establish a strong foundation for lightweight ultrasound AI in practical settings.

VII. Conclusion

In this paper, we proposed TinyUSFM, the first lightweight ultrasound foundation model that addresses the challenges of deploying large-scale foundation models in resource-constrained clinical environments. We developed a feature-gradient driven coreset selection strategy and a domain-separated MIM assisted consistency-driven dynamic distillation to enable effective knowledge transfer while preserving ultrasound-specific characteristics. We also established UniUS-Bench, the largest publicly available ultrasound benchmark comprising 8 classification and 10 segmentation datasets across 15 organs. Experimental results on UniUS-Bench demonstrate that TinyUSFM closely matches the performance of the large-scale USFM while requiring significantly fewer parameters and computational resources, maintaining strong generalization across diverse organs, devices, and clinical centers. This work opens new possibilities for deploying efficient ultrasound AI models in point-of-care settings and resource-limited healthcare environments, potentially improving diagnostic accessibility and clinical workflow efficiency.

References

[1] S. Liu, Y. Wang, X. Yang, B. Lei, L. Liu, S. X. Li, D. Ni, and T. Wang, "Deep learning in medical ultrasound analysis: a review," *Engineering*, vol. 5, no. 2, pp. 261–275, 2019.

[2] Z. Lin, S. Li, S. Wang, Z. Gao, Y. Sun, C.-T. Lam, X. Hu, X. Yang, D. Ni, and T. Tan, "An orchestration learning framework for ultrasound imaging: Prompt-guided hyper-perception and attention-matching downstream synchronization," *Medical Image Analysis*, p. 103639, 2025.

[3] M. Moor, O. Banerjee, Z. S. H. Abad, H. M. Krumholz, J. Leskovec, E. J. Topol, and P. Rajpurkar, "Foundation models for generalist medical artificial intelligence," *Nature*, vol. 616, no. 7956, pp. 259–265, 2023.

[4] J. Jiao, J. Zhou, X. Li, M. Xia, Y. Huang, L. Huang, N. Wang, X. Zhang, S. Zhou, Y. Wang, *et al.*, "Usfm: A universal ultrasound foundation model generalized to tasks and organs towards label efficient image analysis," *Medical image analysis*, vol. 96, p. 103202, 2024.

[5] S. Lu, Y. Chen, Y. Chen, P. Li, J. Sun, C. Zheng, Y. Zou, B. Liang, M. Li, Q. Jin, *et al.*, "General lightweight framework for vision foundation model supporting multi-task and multi-center medical image analysis," *Nature Communications*, vol. 16, no. 1, p. 2097, 2025.

[6] S. Mehta and M. Rastegari, "Mobilevit: light-weight, general-purpose, and mobile-friendly vision transformer," *arXiv preprint arXiv:2110.02178*, 2021.

[7] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International conference on machine learning*, pp. 6105–6114, PMLR, 2019.

[8] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[9] L. J. Brattain, B. A. Telfer, M. Dhyani, J. R. Grajo, and A. E. Samir, "Machine learning for medical ultrasound: status, methods, and future opportunities," *Abdominal radiology*, vol. 43, no. 4, pp. 786–799, 2018.

[10] A. M. Güneş, W. van Rooij, S. Gulshad, B. Slotman, M. Dahele, and W. Verbakel, "Impact of imperfection in medical imaging data on deep learning-based segmentation performance: An experimental study using synthesized data," *Medical physics*, vol. 50, no. 10, pp. 6421–6432, 2023.

[11] R. Compton, L. Zhang, A. Puli, and R. Ranganath, "When more is less: Incorporating additional datasets can hurt performance by introducing spurious correlations," in *Machine learning for healthcare conference*, pp. 110–127, PMLR, 2023.

[12] A. Patra, Y. Cai, P. Chatelain, H. Sharma, L. Drukker, A. T. Papageorghiou, and J. A. Noble, "Efficient ultrasound image analysis models with sonographer gaze assisted distillation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 394–402, Springer, 2019.

[13] J. Qiu, L. Li, J. Sun, J. Peng, P. Shi, R. Zhang, Y. Dong, K. Lam, F. P.-W. Lo, B. Xiao, *et al.*, "Large ai models in health informatics: Applications, challenges, and the future," *IEEE Journal of Biomedical and Health Informatics*, vol. 27, no. 12, pp. 6074–6087, 2023.

[14] G. Wang, J. Wu, X. Luo, X. Liu, K. Li, and S. Zhang, "Mis-fm: 3d medical image segmentation using foundation models pretrained on a large-scale unannotated dataset," *arXiv preprint arXiv:2306.16925*, 2023.

[15] Y. Zhou, M. A. Chia, S. K. Wagner, M. S. Ayhan, D. J. Williamson, R. R. Struyven, T. Liu, M. Xu, M. G. Lozano, P. Woodward-Court, *et al.*, "A foundation model for generalizable disease detection from retinal images," *Nature*, vol. 622, no. 7981, pp. 156–163, 2023.

[16] Z. Wang, C. Liu, S. Zhang, and Q. Dou, "Foundation model for endoscopy video analysis via large-scale self-supervised pre-train," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 101–111, Springer, 2023.

[17] J. Ma, Y. He, F. Li, L. Han, C. You, and B. Wang, "Segment anything in medical images," *Nature Communications*, vol. 15, no. 1, p. 654, 2024.

[18] S. Zhang and D. Metaxas, "On the challenges and perspectives of foundation models for medical image analysis," *Medical image analysis*, vol. 91, p. 102996, 2024.

[19] A. Romero, N. Ballas, S. E. Kahou, A. Chassang, C. Gatta, and Y. Bengio, "Fitnets: Hints for thin deep nets. arxiv 2014," *arXiv preprint arXiv:1412.6550*, 2014.

[20] Y. Liu, K. Chen, C. Liu, Z. Qin, Z. Luo, and J. Wang, "Structured knowledge distillation for semantic segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2604–2613, 2019.

[21] D. Qin, J.-J. Bu, Z. Liu, X. Shen, S. Zhou, J.-J. Gu, Z.-H. Wang, L. Wu, and H.-F. Dai, "Efficient medical image segmentation based on knowledge distillation," *IEEE Transactions on Medical Imaging*, vol. 40, no. 12, pp. 3820–3831, 2021.

[22] Y. Xie, Y. Yin, Q. Li, and Y. Wang, "Deep mutual distillation for semi-supervised medical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 540–550, Springer, 2023.

[23] S. Du, N. Bayasi, G. Hamarneh, and R. Garbi, "Mdvit: Multi-domain vision transformer for small medical image segmentation datasets," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 448–458, Springer, 2023.

[24] W. Zou, X. Qi, W. Zhou, M. Sun, Z. Sun, and C. Shan, "Graph flow: Cross-layer graph flow distillation for dual efficient medical image segmentation," *IEEE Transactions on Medical Imaging*, vol. 42, no. 4, pp. 1159–1171, 2022.

[25] X. Qi, Z. Wu, W. Zou, M. Ren, Y. Gao, M. Sun, S. Zhang, C. Shan, and Z. Sun, "Exploring generalizable distillation for efficient medical image segmentation," *IEEE Journal of Biomedical and Health Informatics*, vol. 28, no. 7, pp. 4170–4183, 2024.

[26] M. Paul, S. Ganguli, and G. K. Dziugaite, "Deep learning on a data diet: Finding important examples early in training," *Advances in neural information processing systems*, vol. 34, pp. 20596–20607, 2021.

[27] B. Mirzasoleiman, J. Bilmes, and J. Leskovec, "Coresets for data-efficient training of machine learning models," in *International Conference on Machine Learning*, pp. 6950–6960, PMLR, 2020.

[28] A. Katharopoulos and F. Fleuret, "Not all samples are created equal: Deep learning with importance sampling," in *International conference on machine learning*, pp. 2525–2534, PMLR, 2018.

[29] P. W. Koh and P. Liang, "Understanding black-box predictions via influence functions," in *International conference on machine learning*, pp. 1885–1894, PMLR, 2017.

[30] Y. Geifman and R. El-Yaniv, "Deep active learning over the long tail," *arXiv preprint arXiv:1711.00941*, 2017.

[31] V. Birodkar, H. Mobahi, and S. Bengio, "Semantic redundancies in image-classification datasets: The 10% you don't need," *arXiv preprint arXiv:1901.11409*, 2019.

[32] G. Holste, Y. Zhou, S. Wang, A. Jaiswal, M. Lin, S. Zhuge, Y. Yang, D. Kim, T.-H. Nguyen-Mau, M.-T. Tran, *et al.*, "Towards long-tailed, multi-label disease classification from chest x-ray: Overview of the cxr-lt challenge," *Medical Image Analysis*, vol. 97, p. 103224, 2024.

[33] K. M. Meiburger, F. Marzola, G. Zahnd, F. Faita, C. P. Loizou, N. Lainé, C. Carvalho, D. A. Steinman, L. Gibello, R. M. Bruno, *et al.*, "Carotid ultrasound boundary study (cubs): Technical considerations on an open multi-center analysis of computerized measurement systems for intima-media thickness measurement on common carotid artery longitudinal b-mode ultrasound scans," *Computers in biology and medicine*, vol. 144, p. 105333, 2022.

[34] P. Cai, T. Yang, Q. Xie, P. Liu, and P. Li, "A lightweight hybrid model for the automatic recognition of uterine fibroid ultrasound images based on deep learning," *Journal of Clinical Ultrasound*, vol. 52, no. 6, pp. 753–762, 2024.

[35] H. Gong, J. Chen, G. Chen, H. Li, G. Li, and F. Chen, "Thyroid region prior guided attention for ultrasound segmentation of thyroid nodules," *Computers in biology and medicine*, vol. 155, p. 106389, 2023.

[36] F. Marzola, N. Van Alfen, J. Doorduin, and K. M. Meiburger, "Deep learning segmentation of transverse musculoskeletal ultrasound images for neuromuscular disease assessment," *Computers in Biology and Medicine*, vol. 135, p. 104623, 2021.

[37] X. Yiming, Z. Bowen, L. Xiaohong, W. Tao, J. Jinxiu, W. Shijie, L. Yufan, Z. Hongjun, L. Tong, S. Ye, J. Rui, W. Guangyu, R. Jie, and C. Ting, "Annotated ultrasound liver images," Nov. 2022.

[38] W. Al-Dhabyani, M. Gomaa, H. Khaled, and A. Fahmy, "Dataset of breast ultrasound images," *Data in brief*, vol. 28, p. 104863, 2020.

[39] Q. Zhao, S. Lyu, W. Bai, L. Cai, B. Liu, M. Wu, X. Sang, M. Yang, and L. Chen, "A multi-modality ovarian tumor ultrasound image dataset for unsupervised cross-domain semantic segmentation," *CoRR*, 2022.

[40] X. P. Burgos-Artizzu, D. Coronado-Gutiérrez, B. Valenzuela-Alcaraz, E. Bonet-Carne, E. Eixarch, F. Crispi, and E. Gratacós, "Evaluation of deep convolutional neural networks for automatic classification of common maternal fetal ultrasound planes," *Scientific Reports*, vol. 10, no. 1, p. 10200, 2020.

[41] C. J. Belasso, B. Behboodi, H. Benali, M. Boily, H. Rivaz, and M. Fortin, "Luminous database: lumbar multifidus muscle segmentation from ultrasound images," *BMC Musculoskeletal Disorders*, vol. 21, no. 1, p. 703, 2020.

[42] R. Singla, C. Ringstrom, G. Hu, V. Lessoway, J. Reid, C. Nguan, and R. Rohling, "The open kidney ultrasound data set," in *International Workshop on Advances in Simplifying Medical Ultrasound*, pp. 155–164, Springer, 2023.

[43] Q. He, S. Bano, J. Liu, W. Liu, D. Stoyanov, and S. Zuo, "Query2: Query over queries for improving gastrointestinal stromal tumour detection in an endoscopic ultrasound," *Computers in Biology and Medicine*, vol. 152, p. 106424, 2023.

[44] L. Pedraza, C. Vargas, F. Narváez, O. Durán, E. Muñoz, and E. Romero, "An open access thyroid ultrasound image database," in *10th International symposium on medical information processing and analysis*, vol. 9287, pp. 188–193, SPIE, 2015.

[45] W. Gómez-Flores, M. J. Gregorio-Calas, and W. Coelho de Albuquerque Pereira, "Bus-bra: a breast ultrasound dataset for assessing computer-aided diagnosis systems," *Medical Physics*, vol. 51, no. 4, pp. 3110–3123, 2024.

[46] A. Montoya, Hasnin, kaggle446, shirzad, W. Cukierski, and yffud, "Ultrasound nerve segmentation." `https://kaggle.com/competitions/ultrasound-nerve-segmentation`, 2016. Kaggle.

[47] J. R. McLaughlan, L. Howell, and N. Ingram, "Lung ultrasound covid phantom dataset used for training machine learning model," 2024.

[48] B. Jieyun and O. ZhanHong, "Pubic symphysis-fetal head segmentation and angle of progression," Apr. 2023.

[49] S. Leclerc, E. Smistad, J. Pedrosa, A. Østvik, F. Cervenansky, F. Espinosa, T. Espeland, E. A. R. Berg, P.-M. Jodoin, T. Grenier, *et al.*, "Deep learning for segmentation using an open large-scale dataset in

2d echocardiography," *IEEE transactions on medical imaging*, vol. 38, no. 9, pp. 2198–2210, 2019.

[50] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[51] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[52] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *International conference on machine learning*, pp. 10347–10357, PMLR, 2021.

[53] Z. Xie, Z. Zhang, Y. Cao, Y. Lin, J. Bao, Z. Yao, Q. Dai, and H. Hu, "Simmim: A simple framework for masked image modeling," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9653–9663, 2022.

[54] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou, "Transunet: Transformers make strong encoders for medical image segmentation," *arXiv preprint arXiv:2102.04306*, 2021.

[55] H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, and M. Wang, "Swin-unet: Unet-like pure transformer for medical image segmentation," in *European conference on computer vision*, pp. 205–218, Springer, 2022.

[56] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "Segformer: Simple and efficient design for semantic segmentation with transformers," *Advances in neural information processing systems*, vol. 34, pp. 12077–12090, 2021.

[57] Q. Wan, Z. Huang, J. Lu, G. Yu, and L. Zhang, "Seaformer: Squeeze-enhanced axial transformer for mobile semantic segmentation," in *The eleventh international conference on learning representations*, 2023.

[58] J. Healy and L. McInnes, "Uniform manifold approximation and projection," *Nature Reviews Methods Primers*, vol. 4, no. 1, p. 82, 2024.