Domain-Constrained Distillation of DINOv3 into a Lightweight Foundation Model toward Point-of-Care Ultrasound

Md Jaber Al Nahian[*1] MDJABER@UALBERTA.CA
Shrimanti Ghosh[*1] SHRIMANT@UALBERTA.CA
Jacob Jaremko[1] JJAREMKO@UALBERTA.CA
Abhilash Hareendranathan[†1] HAREENDR@UALBERTA.CA
[1] *Faculty of Medicine and Dentistry-Radiology and Diagnostic Imaging Department, University of Alberta, AB, CA*

Editors: Under Review for MIDL 2026

Abstract

Vision foundation models such as DINOv3 provide powerful representations but are too computationally demanding for point-of-care ultrasound (POCUS), whereas lightweight CNNs remain deployable yet brittle when faced with diverse anatomies and acquisition styles. We bridge this gap with a domain-constrained distillation framework that transfers DINOv3 ViT-B/16 knowledge into a compact ResNet-50, achieving roughly 3.4× compression while preserving the teacher's billion-scale visual priors. Using a large, heterogeneous ultrasound corpus and physics-aware augmentations, the distilled model delivers substantial linear-probe improvements over standard CNN baselines and consistently outperforms the ViT teacher on challenging, heterogeneous datasets. It further offers marked gains in limited-label regimes, reflecting the realities of POCUS workflows where annotated data are scarce. Embedding visualizations show that the distilled encoder forms clearer, anatomy-aware clusters than the teacher, indicating successful alignment to ultrasound structure. Together, these results demonstrate that large-scale natural-image priors can be distilled into a lightweight, generalizable encoder suitable for resource-constrained clinical deployment.

Keywords: DINOv3, Distillation, POCUS, Foundation Model, Domain Adaptation.

1. Introduction

Large vision foundation models (FMs) such as DINOv3 and related self-supervised ViT encoders achieve strong transfer across many visual tasks by pretraining on billions of natural images (Radford et al., 2021; Oquab et al., 2023; Siméoni et al., 2025; Kirillov et al., 2023). These models are increasingly attractive for medical imaging, where labeled data are scarce and distribution shifts are common. However, most existing FMs are require massive computational resources including GPUs and large memory. These resources are often not available in clinical settings. For instance portable point-of-care ultrasound (POCUS) devices are often connected to tablet or smartphone. Using existing foundation models like DINOv3 on these devices for inference on these devices is challenging and often impractical (Kim et al., 2024).

[*] Corresponding Author
[†] Co-corresponding Author

© 2026 CC-BY 4.0, M.J.A. Nahian, S. Ghosh, J. Jaremko & A. Hareendranathan.

ViT-based models are powerful, yet their high memory and processing requirements make them challenging to run efficiently on POCUS devices (Saha and Xu, 2025). Compact convolutional networks (CNNs) are much easier to deploy, but they are usually trained on small, single-center ultrasound datasets and often fail to generalize across anatomies, scanners, and acquisition protocols (Wu et al., 2024). In practice, clinicians must choose between accurate but impractical models and practical but brittle ones.

We tackle this deployment–performance trade-off by treating FM adaptation as a *knowledge preservation* problem rather than a pure compression problem. Starting from a DINOv3 ViT-B/16 teacher pretrained on 1.7B natural images, we distill its representations into a lightweight ResNet-50 student trained on a curated, large-scale ultrasound corpus of 162,000 unlabeled B-mode images from 40 diverse public datasets. The aim is not only to reduce parameter count, but to transfer both generic visual structure learned at billion scale (edges, shapes, hierarchical abstractions) and ultrasound-specific appearance patterns shaped by B-mode physics and clinical scanning practice. We implement this *domain-constrained* adaptation by supervising the student only through DINOv3 token embeddings, while training on ultrasound-only data with ultrasound-aware augmentations (horizontal flips, moderate zoom, mild blur) that reflect plausible B-mode acquisition changes.

We show that this distillation strategy yields a compact ultrasound foundation model that is competitive with, and in some cases surpasses, the heavy ViT teacher on ultrasound segmentation and classification tasks. It also maintains strong performance in low-label regimes. Representation analyses suggest that the distilled model retains useful natural-image structure while forming anatomically meaningful clusters across diverse ultrasound domains. Overall, our results suggest that billion-scale natural-image pretraining can be transferred into a lightweight CNN without sacrificing accuracy, offering a promising step toward foundation models that better align with the computational constraints of POCUS systems. A detailed discussion of related work on vision FMs, medical FMs, and medical distillation is provided in Section 2.

2. Related Work

2.1. Ultrasound-specific deep learning

Before FM-style pretraining, ultrasound applications primarily relied on task-specific CNNs trained from scratch or initialized from supervised ImageNet weights (Perdios et al., 2018; Zheng et al., 2023; Inan et al., 2024). U-Net variants and ResNet-based encoders have been widely deployed for lesion segmentation, organ boundary detection, and view classification (Ronneberger et al., 2015; Chen et al., 2018). While compact enough for embedded deployment, these models are typically trained on small, single-center datasets and exhibit poor cross-domain generalization.

2.2. Vision foundation models and knowledge distillation

Large-scale vision foundation models including CLIP, DINOv2/DINOv3, and Segment Anything (SAM/SAM2) achieve strong zero-shot and transfer performance across classification, detection, and segmentation by pretraining on hundreds of millions to billions of natural images (Radford et al., 2021; Oquab et al., 2023; Siméoni et al., 2025; Kirillov et al., 2023).

Knowledge distillation—where a large teacher supervises a smaller student via feature, logit, or attention matching—is widely used to compress such models for edge deployment (Hinton et al., 2015; Romero et al., 2015; Zagoruyko and Komodakis, 2016). Recent work has distilled SAM-like segmenters and DINO-style self-supervised ViTs into compact students, showing that much of the teacher's representational power can be retained in lighter architectures (Zhang et al., 2023; Kang et al., 2023). These approaches, however, are almost exclusively evaluated on natural-image benchmarks.

2.3. Foundation models in medical imaging

Medical imaging has rapidly adopted foundation-model pretraining, with adaptations of SAM (e.g., MedSAM, Sam2Rad) and other large encoders improving performance and label efficiency across CT, MRI, X-ray, and histopathology (Ma et al., 2024; Wahd et al., 2025; Hosseinzadeh Taher et al., 2023; Pai et al., 2025; Shaikovski et al., 2024). For ultrasound specifically, emerging ultrasound foundation models trained across multiple organs and anatomies demonstrate promising transfer to segmentation and classification under limited labels (Megahed et al., 2025; Jiao et al., 2024; Ma et al., 2025). Most of these models, however, retain heavy ViT or large CNN backbones and assume access to data-center–class hardware at inference time.

2.4. Knowledge distillation in medical imaging

Knowledge distillation has been applied in medical imaging to compress large segmentation networks, ensembles, and self-supervised encoders (Qin et al., 2021; Wang et al., 2023; Vray et al., 2024). Existing techniques distill logits, intermediate feature maps, or contrastive representations, sometimes with uncertainty weighting or region-aware losses. Most prior work either uses supervised task-specific teachers (discarding a wealth of natural-image priors), focuses on a single modality or anatomy, or treats domain adaptation as downstream fine-tuning rather than an integral part of the distillation process. In ultrasound, distillation has mainly been used to compress task-specific models rather than to build general-purpose ultrasound foundation models (Dapueto et al., 2024).

2.5. Main Contribution

Overall, existing foundation models and distillation methods do not systematically address how to preserve large-scale natural-image priors while adapting to specialized medical modalities such as ultrasound. In practice, current ultrasound approaches still trade off between heavy ViT-based FMs (high capacity but impractical for point-of-care deployment) and lightweight CNNs (easy to deploy but not generalizable). Motivated by this gap, the next section introduces a domain-constrained, feature-level distillation framework that transfers DINOv3 token representations into a compact ResNet-50 student, using ultrasound-only data and ultrasound-aware augmentations to tailor the encoder for POCUS settings.

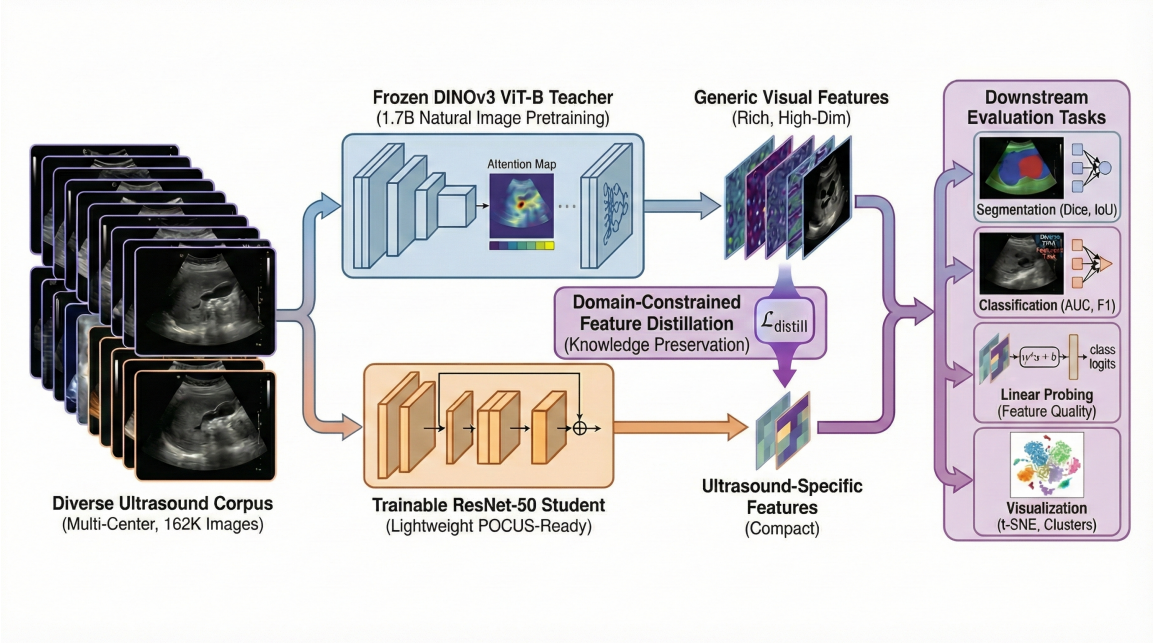

Figure 1: Overview of the proposed domain-constrained distillation framework. A frozen DINOv3 ViT-B/16 teacher, pretrained on 1.7B natural images, produces rich generic visual features. We distill these features into a lightweight ResNet-50 student using a diverse multi-center ultrasound corpus (162K images) and ultrasound-aware augmentations, optimizing a token-wise feature loss $\mathcal{L}_{\text{distill}}$ for knowledge preservation. The resulting compact encoder yields ultrasound-specific features that are evaluated on downstream segmentation and classification tasks, as well as via linear probing and representation visualization.

3. Method

Our goal is to obtain a compact, ultrasound-specific encoder by distilling a large DINOv3 ViT-B/16 teacher into a ResNet-50 student using a large corpus of unlabeled ultrasound images. Fig. 1 summarizes the pipeline. In this section, we describe the ultrasound corpus, the teacher–student architecture, the feature-level distillation objective, the ultrasound-aware augmentations, and the optimization details.

3.1. Ultrasound Corpus

We curate a heterogeneous corpus of ∼160,000 unlabeled B-mode ultrasound images spanning diverse anatomies (breast, thyroid, cardiac, fetal, musculoskeletal) and acquisition settings. Images are extracted from public and institutional datasets, standardized to a fixed resolution, and cleaned by removing overlays and discarding non-ultrasound or corrupted frames. No labels are used during training, encouraging the encoder to learn transferable, anatomy-agnostic ultrasound representations.

3.2. Teacher and student architectures

We adopt a high-capacity DINOv3 vision transformer as the teacher and a compact ResNet as the student.

Teacher: DINOv3 ViT-B/16. The teacher encoder T is a DINOv3 ViT-B/16 model pre-trained self-supervised on $\approx$ 1.7 billion natural images. For an input image $x \in \mathbb{R}^{3\times H\times W}$, the ViT processes non-overlapping 16×16 patches and produces a sequence of token embeddings

$$Z_T(x) \in \mathbb{R}^{N_{\text{tok}}\times d_T}, \tag{1}$$

where N_{tok} is the number of image tokens and d_T is the teacher embedding dimension. In our setup, token features from n intermediate transformer blocks (we set $n = 2$) are extracted and aggregated to form the teacher representation used for distillation.

Student: ResNet-50. The student encoder S is a standard ResNet-50. Given x, the student produces a convolutional feature map

$$F_S(x) \in \mathbb{R}^{C_S\times H_S\times W_S}. \tag{2}$$

This feature map is reshaped into a sequence of spatial tokens and passed through a small projection head g_ϕ (an MLP with two layers and a hidden dimension of 4096) to match the teacher embedding dimension:

$$Z_S(x) = g_\phi\big(\text{flatten}(F_S(x))\big) \in \mathbb{R}^{N_{\text{tok}}\times d_T}. \tag{3}$$

The ViT-B/16 teacher has $\sim$ 86M parameters, whereas the ResNet-50 student has $\sim$ 25M parameters, yielding a $\sim$ 3–4$\times$ reduction in parameter count and model size.

3.3. Feature-level distillation

We use a feature-level knowledge distillation scheme that aligns teacher and student token embeddings on unlabeled ultrasound images. For each image x, we apply a stochastic augmentation $a(\cdot)$ to obtain $\tilde{x} = a(x)$ and feed the same view to both teacher and student:

$$Z_T = Z_T(\tilde{x}) = T(\tilde{x}) \in \mathbb{R}^{N_{\text{tok}}\times d_T}, \tag{4}$$

$$Z_S = Z_S(\tilde{x}) = g_\phi\big(\text{flatten}(F_S(\tilde{x}))\big) \in \mathbb{R}^{N_{\text{tok}}\times d_T}. \tag{5}$$

Token Alignment. To ensure spatial consistency during distillation, we align the ViT-B/16 teacher's 16×16 token grid with the ResNet-50 student's output. We extract n intermediate feature maps from the teacher, resize them if necessary, and concatenate them into a unified token sequence. The student's 8×8 feature map is projected via an MLP and upsampled to match the teacher's resolution. Both outputs are reshaped into sequences and supervised using a mean squared error loss, enabling effective token-wise alignment without modifying either backbone.

Distillation loss. The distillation loss is implemented as a token-wise mean squared error (MSE) between teacher and student embeddings. Let $Z_T(i) \in \mathbb{R}^{d_T}$ and $Z_S(i) \in \mathbb{R}^{d_T}$ denote the embeddings of the i-th token in the sequence. The loss for one image is

$$\mathcal{L}_{\text{distill}}(x) = \frac{1}{N_{\text{tok}}} \sum_{i=1}^{N_{\text{tok}}} \left\| Z_T(i) - Z_S(i) \right\|_2^2, \tag{6}$$

and the batch loss is obtained by averaging (6) across the mini-batch. The loss returns a single scalar and is exactly zero when teacher and student features are identical. We do not use any additional logit-based distillation or contrastive loss; all supervision is mediated through the teacher token embeddings.

Mixup regularization. We further apply image-level mixup within each batch as a regularizer. Given two images x_a and x_b and a mixing coefficient $\lambda \sim \mathcal{U}(0, 1)$, the mixed input is

$$\tilde{x}_{\text{mix}} = \lambda \tilde{x}_a + (1 - \lambda)\tilde{x}_b, \tag{7}$$

and the corresponding teacher and student features are linearly interpolated. This encourages smoother transitions in feature space and improves stability during training with large batches.

3.4. Ultrasound-aware data augmentations

We restrict distillation to transformations that reflect real ultrasound acquisition. Images are resized and cropped to mimic natural variation in zoom and field-of-view; horizontal flips are allowed, but vertical flips are excluded because ultrasound probes have a fixed orientation relative to the skin surface, making upside-down views physically impossible in clinical practice. Mild Gaussian blur models depth-dependent defocus, and color jitter is removed because ultrasound is inherently grayscale. These choices ensure that both teacher and student learn invariances tied to actual probe motion and imaging physics, forming a strictly "domain-constrained" augmentation pipeline.

3.5. Optimization and implementation details

We distill a ResNet-50 student from a DINOv3 ViT-B/16 teacher using mini-batch training on the unlabeled ultrasound corpus. A lightweight two-layer MLP projects student features into the teacher embedding space, and tokens from two intermediate ViT blocks are used as supervision. Training runs for 1000 epochs with AdamW (learning rate 1×10^{-4}, weight decay 0.05), batch size 512, and mixed bfloat16 precision on NVIDIA GPUs. Checkpoints are saved periodically and training can resume after interruptions.

4. Experiments

4.1. Tasks and datasets

We evaluate the proposed encoder on two segmentation tasks and one classification task. DDTI is a thyroid nodule segmentation dataset of 637 B-mode ultrasound images with expert pixel-level nodule masks (Pedraza et al., 2015). We use 445 images for training,

127 for validation, and 65 for testing, with splits constructed at the patient level to avoid leakage. BUSI (Al-Dhabyani et al., 2020) is a breast ultrasound dataset of 780 images with lesion masks and image-level labels. For BUSI segmentation, we use the provided binary lesion masks for benign, malignant, and normal cases and create train, validation, and test splits analogous to DDTI. For BUSI classification, we define a three-class problem with normal (133 images), benign (437 images), and malignant (210 images), using disjoint class-stratified train, validation, and test sets built from class-wise folders.

4.2. Models and baselines

All downstream experiments are based on ResNet-50 backbones under three initialization schemes: (i) **R50-Rand**, a ResNet-50 trained from scratch; (ii) **R50-Distill-Default** (ours), a ResNet-50 obtained by distilling a DINOv3 ViT-B/16 teacher using default natural-image augmentations; and (iii) **R50-Distill-US** (ours), a ResNet-50 obtained by distilling the same teacher on an ultrasound-only corpus with ultrasound-aware augmentations as described in Section 3. For segmentation, each ResNet-50 variant is used as the encoder in the same U-Net-style architecture implemented with Segmentation Models PyTorch, so that only the backbone initialization differs. As a high-capacity reference, we also fine-tune a DINOv3 ViT-B/16 backbone with a lightweight segmentation head. For BUSI classification, we use a standard ResNet-50 classifier (global average pooling followed by a linear head) with the three initializations above.

4.3. Training protocols and evaluation metrics

For DDTI and BUSI segmentation, all models use the same U-Net decoder with a ResNet-50 or ViT-B/16 encoder at a fixed input resolution. Training uses standard geometric augmentations, while validation/test images undergo only resizing and normalization. Linear probing freezes the encoder to isolate representation quality, and full fine-tuning updates all parameters under identical optimization settings. For BUSI classification, we fine-tune a ResNet-50 with a three-way output head using the same preprocessing pipeline. All initialization variants (R50-Rand, R50-Distill-Default, R50-Distill-US) share identical training schedules, and model selection is based on validation accuracy. For segmentation on DDTI and BUSI, we report mean Dice coefficient and mean Intersection-over-Union (mIoU) on the held-out test sets. For BUSI classification, we report overall test accuracy and macro-averaged F1 score.

5. Results

5.1. Linear probing on frozen encoders

Table 1 summarizes linear-probe performance across segmentation and classification tasks. On DDTI, both distilled models substantially outperform the randomly initialized baseline, with **R50-Distill-Default** obtaining the best Dice (0.7378) and IoU (0.6252), indicating successful transfer of ViT teacher knowledge to a compact CNN. In contrast, the **DINOv3 ViT-B/16** teacher underperforms (Dice 0.6503), reflecting limited robustness to thyroid-domain grayscale and speckle statistics. On BUSI, the domain gap becomes more

Table 1: Linear probing on frozen encoders. DDTI and BUSI segmentation are evaluated by mean Dice and mean IoU. BUSI classification is a 3-way task (normal/benign/malignant) evaluated by accuracy and macro F1.

Model	DDTI Seg.		BUSI Seg.		BUSI Cls.	
	Dice	IoU	Dice	IoU	Acc	F1
R50-Rand	0.6028	0.4647	0.4365	0.3369	0.6115	0.4204
R50-Distill-Default	**0.7378**	**0.6252**	0.5771	0.4857	0.7325	**0.7037**
R50-Distill-US	0.7334	0.6192	**0.6083**	**0.5259**	**0.7452**	0.6768
DINOv3 ViT-B/16	0.6503	0.4743	0.2384	0.1729	–	–

Table 2: Full fine-tuning on all labeled data. Metrics as in Table 1.

Model	DDTI Seg.		BUSI Seg.		BUSI Cls.	
	Dice	IoU	Dice	IoU	Acc	F1
R50-Rand	0.6605	0.5297	0.5525	0.4629	0.6561	0.6526
R50-Distill-Default	0.7652	0.6608	**0.6967**	**0.6209**	0.8662	0.8572
R50-Distill-US	0.7872	**0.6745**	0.6930	0.6126	**0.8790**	**0.8673**
DINOv3 ViT-B/16	**0.7933**	0.4790	0.2838	0.1957	–	–

pronounced: the teacher collapses to a Dice of 0.2384, while the ultrasound-aware **R50-Distill-US** achieves the strongest segmentation performance (Dice 0.6083, IoU 0.5259). Similarly, **R50-Distill-US** yields the best BUSI classification accuracy (0.7452), whereas **R50-Distill-Default** attains the highest macro F1 (0.7037). These results confirm that distillation on ultrasound-only data, paired with physics-consistent augmentations, produces representations substantially better aligned with downstream ultrasound tasks than the generic natural-image ViT teacher.

5.2. Full fine-tuning on all labeled data

As shown in Table 2, full fine-tuning substantially amplifies the benefits of distillation. Both distilled ResNet-50 models outperform the randomly initialized baseline across all tasks, and the ultrasound-aware variant closely matches the ViT teacher on DDTI despite having only a fraction of the parameters. The ViT's inconsistent Dice–IoU behavior further suggests difficulty aligning its natural-image features with ultrasound boundary structure. On BUSI, the contrast is even stronger: the teacher fails to adapt, while both distilled students fine-tune reliably and achieve markedly better segmentation and classification performance. This indicates that distillation not only compresses the teacher but also removes natural-image biases that hinder direct transfer to ultrasound.

5.3. Limited-label regimes

As shown in Table 3, both distilled models maintain strong performance even when fine-tuned with only a small fraction of labeled data, whereas the randomly initialized baseline degrades quickly. The ultrasound-aware student is particularly stable in the lowest-label settings, indicating that domain-constrained distillation yields features that transfer more reliably under scarce supervision. These trends are most evident in BUSI classification,

Table 3: Limited-data performance for different label fractions. We report segmentation Dice on DDTI (thyroid) and BUSI (breast), and BUSI 3-way classification accuracy, when fine-tuning on 5%, 10%, 20%, and 50% of labeled data.

	DDTI Dice				BUSI Dice				BUSI Acc			
Model	5%	10%	20%	50%	5%	10%	20%	50%	5%	10%	20%	50%
R50-Rand	0.4269	0.5237	0.5597	0.6007	0.1911	0.3381	0.3895	0.4038	0.2675	0.5605	0.6242	0.6561
R50-Distill-Default	0.4958	0.6536	0.7109	**0.7672**	0.2494	**0.4857**	0.5011	**0.6185**	0.4968	0.5860	0.6688	0.7452
R50-Distill-US	**0.5233**	**0.6574**	**0.7309**	0.7661	**0.2752**	0.4600	**0.5036**	0.5543	**0.7197**	**0.6242**	**0.8025**	**0.8726**

Table 4: Model capacity vs. performance after full fine-tuning. Parameter counts are approximate.

Model	Params (M)	DDTI Dice	BUSI Dice	BUSI Acc
DINOv3 ViT-B/16	86	**0.7933**	0.2838	–
R50-Rand	25	0.6605	0.5525	0.6561
R50-Distill-US	25	.7872	0.6930	**0.8790**

where the distilled encoder consistently outperforms alternatives across all label fractions. This highlights a key advantage of our approach: by embedding ultrasound-specific priors during distillation, the model becomes far less dependent on large annotated datasets. Such label efficiency is essential for point-of-care and resource-limited environments, where expert annotation is costly or unavailable.

5.4. Capacity–performance trade-off

Table 4 highlights the deployment benefits of distillation. The ViT teacher is more than three times larger than the ResNet-50 student, yet the distilled model retains similar performance on DDTI and delivers substantially stronger results on BUSI. This indicates that distillation not only compresses the teacher but also yields representations better aligned with ultrasound, making the compact student a more practical choice for point-of-care deployment.

5.5. Representation analysis

t-SNE projections of the ultrasound corpus (Fig. 2) show clear differences in how each model organizes the data. The randomly initialized encoder produces highly entangled embeddings with little anatomical separation, consistent with its weaker downstream performance. Distillation markedly improves structure: the default student forms more coherent clusters, while the ultrasound-aware student produces the most distinct and stable separation across datasets. In contrast, the ViT teacher retains broad natural-image structure but shows substantial overlap between ultrasound domains, mirroring its poor BUSI performance. These patterns support our central claim that domain-constrained distillation realigns feature space toward ultrasound-specific cues, enabling stronger generalization on heterogeneous clinical data.

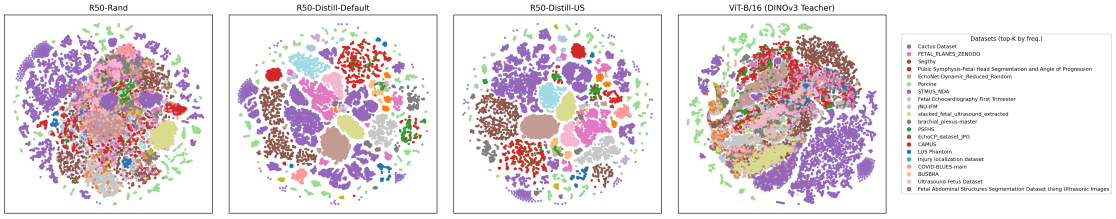

Figure 2: t-SNE visualization of Ultrasound corpus embeddings for four models: R50-Rand, R50-Distill-Default, R50-Distill-US, and ViT-B/16 (DINOv3 Teacher). Points are colored by dataset, with a shared color map across models; the right-most panel shows the legend for the top-K most frequent datasets. R50-Distill-US forms the most compact and well-separated clusters across ultrasound domains.

6. Conclusion

We introduced a domain-constrained distillation framework that transfers billion-scale ViT representations into a compact ResNet-50 suitable for ultrasound. Using a curated ultrasound corpus and ultrasound-aware augmentations, the distilled models offer stronger generalization and substantially better label efficiency than both a randomly initialized CNN and the original DINOv3 ViT teacher. These findings demonstrate that large vision priors can be preserved in a deployment-friendly backbone while mitigating the natural-image biases that limit direct ViT adaptation. Our approach provides a simple and practical recipe for building ultrasound foundation models that are compatible with point-of-care constraints. Limitations include the focus on 2D B-mode data and the absence of on-device latency evaluation. Future work will extend this framework to video-based POCUS, additional anatomies, and hardware-aware model design.

References

Walid Al-Dhabyani, Mohammed Gomaa, Hussien Khaled, and Aly Fahmy. Dataset of breast ultrasound images. *Data in brief*, 28:104863, 2020.

Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-Decoder with atrous separable convolution for semantic image segmentation. *European Conference on Computer Vision*, 2018.

Jacopo Dapueto, Luca Zini, and Francesca Odone. Knowledge distillation for efficient standard scanplane detection of fetal ultrasound. *Medical & Biological Engineering & Computing*, 62(1):73–82, 2024.

Geoffrey Hinton, Vincent Vanhoucke, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Mohammad Reza Hosseinzadeh Taher, Michael B Gotway, and Jianming Liang. Towards foundation models learned from anatomy in medical imaging via self-supervision. In

MICCAI Workshop on Domain Adaptation and Representation Transfer, pages 94–104. Springer, 2023.

Neslihan Gökmen Inan, Ozan Kocadağlı, Düzgün Yıldırım, İsmail Meşe, and Özge Kovan. Multi-class classification of thyroid nodules from automatic segmented ultrasound images: Hybrid resnet based unet convolutional neural network approach. *Computer Methods and Programs in Biomedicine*, 243:107921, 2024.

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.

Dahyun Kang, Piotr Koniusz, Minsu Cho, and Naila Murray. Distilling self-supervised vision transformers for weakly-supervised few-shot classification & segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19627–19638, 2023.

Seungjun Kim, Chanel Fischetti, Megan Guy, Edmund Hsu, John Fox, and Sean D Young. Artificial intelligence (ai) applications for point of care ultrasound (pocus) in low-resource settings: a scoping review. *Diagnostics*, 14(15):1669, 2024.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023.

Chen Ma, Jing Jiao, Shuyu Liang, Junhu Fu, Qin Wang, Zeju Li, Yuanyuan Wang, and Yi Guo. Tinyusfm: Towards compact and efficient ultrasound foundation models. *arXiv preprint arXiv:2510.19239*, 2025.

Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images. *Nature Communications*, 15(1):654, 2024.

Youssef Megahed, Robin Ducharme, Mark Walker, Steven Hawken, and Adrian DC Chan. Usf-mae: Ultrasound self-supervised foundation model with masked autoencoding. *arXiv preprint arXiv:2510.22990*, 2025.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. DINOv2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Suraj Pai, Ibrahim Hadzic, Dennis Bontempi, Keno Bressem, Benjamin H Kann, Andriy Fedorov, Raymond H Mak, and Hugo JWL Aerts. Vision foundation models for computed tomography. *arXiv preprint arXiv:2501.09001*, 2025.

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.

Dimitris Perdios, Manuel Vonlanthen, Adrien Besson, Florian Martinez, Marcel Arditi, and Jean-Philippe Thiran. Deep convolutional neural network for ultrasound image enhancement. In *2018 IEEE International Ultrasonics Symposium (IUS)*, pages 1–4. IEEE, 2018.

Dian Qin, Jia-Jun Bu, Zhe Liu, Xin Shen, Sheng Zhou, Jing-Jun Gu, Zhi-Hua Wang, Lei Wu, and Hui-Fen Dai. Efficient medical image segmentation based on knowledge distillation. *IEEE Transactions on Medical Imaging*, 40(12):3820–3831, 2021.

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.

Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets, 2015. URL https://arxiv.org/abs/1412.6550.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. *International Conference on Medical Image Computing and Computer Assisted Intervention*, pages 234–241, 2015.

Shaibal Saha and Lanyu Xu. Vision transformers on the edge: A comprehensive survey of model compression and acceleration strategies. *Neurocomputing*, page 130417, 2025.

George Shaikovski, Adam Casson, Kristen Severson, Eric Zimmermann, Yi Kan Wang, Jeremy D Kunz, Juan A Retamero, Gerard Oakley, David Klimstra, Christopher Kanan, et al. Prism: A multi-modal generative foundation model for slide-level histopathology. *arXiv preprint arXiv:2405.10254*, 2024.

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. DINOv3. *arXiv preprint arXiv:2508.10104*, 2025.

Guillaume Vray, Devavrat Tomar, Behzad Bozorgtabar, and Jean-Philippe Thiran. Distill-soda: Distilling self-supervised vision transformer for source-free open-set domain adaptation in computational pathology. *IEEE transactions on medical imaging*, 43(5):2021–2032, 2024.

Assefa Seyoum Wahd, Banafshe Felfeliyan, Yuyue Zhou, Shrimanti Ghosh, Adam McArthur, Jiechen Zhang, Jacob L Jaremko, and Abhilash Hareendranathan. Sam2rad: A segmentation model for medical images with learnable prompts. *Computers in Biology and Medicine*, 187:109725, 2025.

Jiping Wang, Yufei Tang, Zhongyi Wu, Qiang Du, Libing Yao, Xiaodong Yang, Ming Li, and Jian Zheng. A self-supervised guided knowledge distillation framework for unpaired low-dose ct image denoising. *Computerized medical imaging and graphics*, 107:102237, 2023.

Derek Wu, Delaney Smith, Blake VanBerlo, Amir Roshankar, Hoseok Lee, Brian Li, Faraz Ali, Marwan Rahman, John Basmaji, Jared Tschirhart, et al. Improving the generalizability and performance of an ultrasound deep learning model using limited multicenter data for lung sliding artifact identification. *Diagnostics*, 14(11):1081, 2024.

Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. *arXiv preprint arXiv:1612.03928*, 2016.

Chaoning Zhang, Dongshen Han, Yu Qiao, Jung Uk Kim, Sung-Ho Bae, Seungkyu Lee, and Choong Seon Hong. Faster segment anything: Towards lightweight sam for mobile applications. *arXiv preprint arXiv:2306.14289*, 2023.

Tianlei Zheng, Hang Qin, Yingying Cui, Rong Wang, Weiguo Zhao, Shijin Zhang, Shi Geng, and Lei Zhao. Segmentation of thyroid glands and nodules in ultrasound images using the improved u-net architecture. *BMC Medical Imaging*, 23(1):56, 2023.