Anatomy-aware Representation Learning for Medical Ultrasound

SeokHwan Oh, Myeong-Gee Kim, Guil Jung, Hyeon-Jik Lee, YoungMin Kim, Sang-Yun Kim, Hyuksool Kwon, Hyeon-Min Bae

INTRODUCTION

- **Medical ultrasound (US)** is widely used medical imaging modality, due to its **non-ionizing, cost-effective, and real-time nature.**
- Medical ultrasound **exhibits limited image quality and high variability**, leading to strong **operator dependence** and motivating the development of computer-assisted diagnostic systems for improved diagnosis.
- Self-supervised learning for medical ultrasound remains relatively underexplored, highlighting the need for **ultrasound-specific representation learning** that account for its distinct characteristics of ultrasound

AIMs

- Configuring one of the **largest-scale datasets dedicated to medical US,** which serves as the foundation for the proposed representation learning
- Developing a robust **US foundation representation learning framework** that generalizes across diverse anatomical domains, captures organ-specific characteristics through anatomy-aware representations

Medical ultrasound image

- Medical ultrasound exhibits **distinct feature from natural image (NI)**

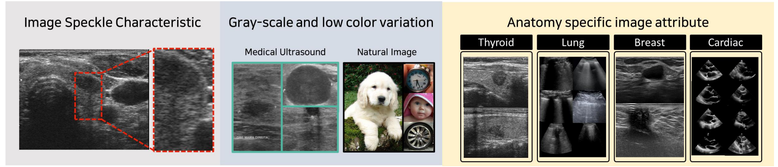
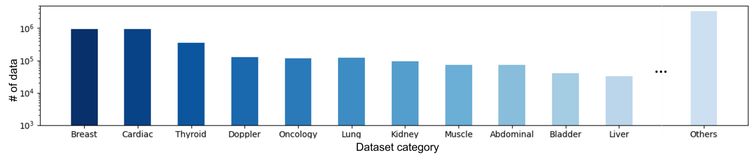


Figure. Difference between medical US and natural images.

Method: Large-scale medical ultrasound dataset formation

- In this paper, we introduce a **large-scale US dataset** comprising **5.2 million images**, covering **16 anatomical categories** and diverse imaging conditions.



Method: Anatomy-Aware Representation Learning

Anatomy-aware Vision Transformer (A-ViT)

- We propose **A-ViT, designed to provide anatomy-adaptive SSL for US images** by incorporating anatomical context using anatomy-conditioned deformable transformer.
- **Anatomy-conditioned deformable transformer (ACDT).** ACDT models **anatomy-dependent spatial feature** distributions by conditioning **deformable convolution** on anatomical context, enabling adaptive receptive fields across different organs.
- The ACDT features are integrated into transformer attention, providing anatomically adaptive representation learning.

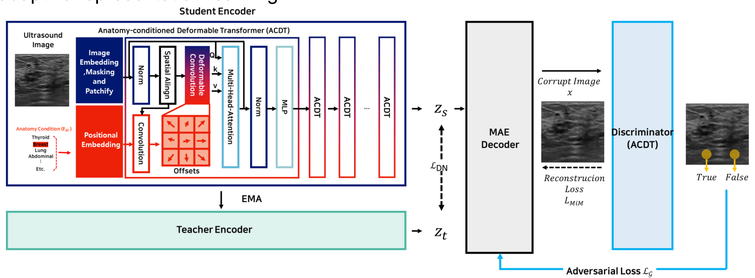


Figure. Overall configuration of the proposed ViT

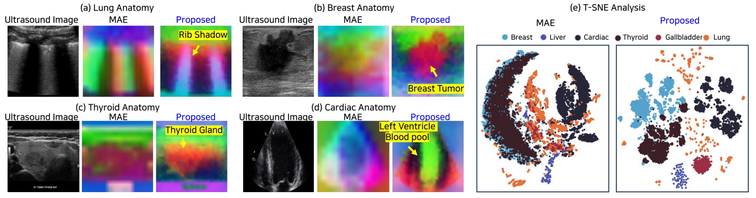
Learning Objective

- To optimize the A-ViT for US image representation, we formulate **an SSL strategy that integrates multiple complementary objectives**.
- The proposed SSL framework combines **masked image modeling (MIM)** for local structural recovery**, an adversarial objective** to preserve high-frequency speckle features **and a self-distillation objective** inspired by DINO for global semantic alignment.

Experiments and results

Qualitative assessment: Feature PCA and t-SNE Analysis

Qualitative assessment demonstrates that **A-ViT learns more discriminative and anatomically faithful representations**, leading to improved performance across diverse ultrasound tasks.



Experiments and results

We have evaluated downstream performance in **major medical US applications** including **breast cancer, thyroid cancer, gallbladder tumor, cardiac view classification, and COVID-19.**

Table. Quantitative assessment of breast cancer classification.

Method	Linear Probing				Fine-tuning			
	Accuracy	AUROC [CI]	Sensitivity	Specificity	Accuracy	AUROC [CI]	Sensitivity	Specificity
Supervised	–	–	–	–	80.98	0.8653 (0.79, 0.92)	0.6667	0.9010
MAE	77.64	0.8211 (0.75, 0.89)	0.5747	0.9000	83.09	0.8635 (0.80, 0.92)	0.7636	0.8778
MoCo v3	83.09	0.9143 (0.86, 0.96)	0.6875	0.9239	85.91	0.9065 (0.85, 0.95)	0.6897	**0.9674**
iBOT	84.50	0.9090 (0.86, 0.95)	0.7632	0.9023	88.02	0.9256 (0.88, 0.96)	0.8347	0.9140
SigLIP2	77.64	0.8396 (0.77, 0.90)	0.6829	0.8352	89.34	0.9351 (0.89, 0.97)	0.8200	0.9468
Dino v3	78.16	0.8382 (0.77, 0.90)	0.6818	0.8446	84.50	0.9172 (0.87, 0.96)	0.6667	0.9326
LVM-Med	82.39	0.8694 (0.81, 0.92)	0.6304	**0.9462**	84.50	0.9083 (0.86, 0.95)	0.8000	0.8790
DMAE	78.16	0.8306 (0.75, 0.90)	0.7188	0.8118	86.61	0.9308 (0.89, 0.97)	0.7238	0.9565
USFM	82.39	0.8927 (0.84, 0.94)	0.7091	0.9022	88.73	0.9376 (0.89, 0.97)	0.8163	0.9344
Proposed	**86.62**	**0.9151 (0.86, 0.96)**	**0.8333**	0.8917	**93.66**	**0.9742 (0.95, 0.99)**	**0.9455**	0.9355

Comparative study: A-ViT consistently **outperforms SoTA SSL baselines in both linear probing and fine-tuning,** demonstrating effective representation learning for medical ultrasound

Table. Ablation study on the effect of loss function and ACDT

$\mathcal{L}_{SD}$	$\mathcal{L}_{MIM}$	$\mathcal{L}_{adv}$	ACDT	Dataset	Accuracy	AUROC [CI]	Sensitivity	Specificity	Params
	✓			Natural Image	83.09	0.8635 (0.80,0.92)	0.7636	0.8778	86M
	✓			Ultrasound	89.43 +6.34	0.9380 (0.89,0.97)	0.8750	0.9140	86M
	✓		✓	Ultrasound	92.25 +2.82	0.9664 (0.93,0.99)	0.8727	**0.9570**	95M
	✓	✓	✓	Ultrasound	92.95 +0.70	0.9688 (0.94,0.99)	0.8820	0.9464	95M
✓	✓	✓	✓	Ultrasound	**93.66** +0.71	**0.9742** (0.95,0.99)	**0.9455**	0.9355	95M

Ablation study: The accuracy is **achieved by ultrasound-specific pretraining, ACDT, and complementary SSL objectives** that enhance both local and global feature learning.

Extended evaluation of medical ultrasound image analysis

Experimental **results demonstrate the effectiveness of ARL over SoTA baselines** in major medical US applications, achieving significant improvements in accuracy and AUROC.

Table. Quantitative assessments on diverse down-stream applications

Method	Cardiac Seg.		Cardiac view Cls.		Thyroid cancer Cls.		COVID-19 Cls.		Gallbladder Tumor Cls.	
	Dice	mIoU	Top-1	Top-3	Acc.	AUROC [CI]	Acc.	AUROC [CI]	Acc.	AUROC [CI]
MAE	89.21	81.06	89.07	99.09	82.50	0.9110 (0.89, 0.93)	80.74	0.9346 (0.92, 0.95)	83.39	0.9105 (0.88, 0.94)
MoCo v3	89.80	82.01	91.08	99.02	83.50	0.9112 (0.89, 0.93)	82.22	0.9286 (0.92, 0.94)	84.47	0.9237 (0.89, 0.95)
SigLIP2	90.60	83.23	90.97	99.11	85.69	0.9330 (0.91, 0.95)	78.05	0.8897 (0.88, 0.90)	84.11	0.9123 (0.88, 0.94)
Dino v3	90.76	83.35	89.93	99.12	86.24	0.9428 (0.93, 0.96)	86.94	0.9465 (0.93, 0.96)	84.84	0.9189 (0.89, 0.95)
iBOT	89.45	81.37	90.59	99.13	83.68	0.9361 (0.92, 0.95)	81.55	0.9494 (0.94, 0.96)	85.19	0.9213 (0.89, 0.95)
LVM-Med	89.41	81.30	88.64	98.72	83.42	0.9101 (0.89, 0.93)	85.40	0.9359 (0.92, 0.95)	80.15	0.8700 (0.82, 0.91)
DMAE	90.44	83.02	90.50	99.20	84.72	0.9210 (0.90, 0.94)	85.99	0.9059 (0.89, 0.92)	84.83	0.9019 (0.87, 0.94)
USFM	91.13	84.15	89.95	98.97	85.50	0.9301 (0.91, 0.95)	87.67	0.9475 (0.94, 0.96)	86.64	0.9347 (0.90, 0.96)
Proposed	**92.16**	**85.67**	**91.80**	**99.22**	**87.07**	**0.9475 (0.93, 0.96)**	**91.44**	**0.9714 (0.97, 0.98)**	**89.89**	**0.9511 (0.93, 0.97)**

Conclusions

- By incorporating the A-ViT and leveraging a newly configured large-scale US dataset, **A-ViT enables generalizable feature extraction adaptive to specific anatomical contexts**
- **Experimental results demonstrate the effectiveness of A-Vit** over state-of-the-art baselines in major medical US applications
- The experiments demonstrate **the potential of the A-Vit as a reliable and efficient solution for computer-aided diagnostics**

Contact Info.

Presenter:
Seokhwan Oh

Email:
shoh@barreleye.co.kr

Project Website:
http://nais.kaist.ac.kr/
http://barreleye.co.kr